Chapter 2
Metric Spaces

Here we give the elements of the theory of metric spaces: the ideas developed in this chapter will be extensively used in the rest of the book.

A metric space is simply a non-empty set X such that to each $x, y \in X$ there corresponds a non-negative number called the distance between x and y. To make the theory sufficiently rich this distance is supposed to have certain properties, such as symmetry and the triangle inequality, that are familiar from Euclidean geometry. As we shall see, the previous chapter offers many examples of such spaces. The idea of a metric space was introduced in 1906 by Fréchet and was significantly developed further in 1914 by Hausdorff, who introduced the term 'metric space'. Further impetus was provided from 1920 onwards by the fundamental work of the Polish school led by Banach: this was largely concerned with the case in which X was a linear space and was of great significance in the establishment of functional analysis as an important part of mathematics. Here we shall not assume that X has any linear structure as neither the results given nor the applications to complex analysis made in the next chapter need this property.

In this chapter we introduce some basic terminology and discuss in detail the fundamental properties of completeness, compactness and connectedness which such spaces may possess; further, special attention is paid to various forms of homotopy and to simple-connectedness. These properties not only have intrinsic interest but also are essential for later work surrounding such central results of complex analysis as, for example, the general version of the famous theorem due to Cauchy. Quite apart from the elegance of metric space theory, it is remarkably useful in that often a single theorem may be applied to handle seemingly different problems. Applications include a proof of the existence of a continuous, nowhere differentiable function, justification of differentiation under the integral sign, and establishment of a solution of an initial-value problem for a certain type of differential equation.

R. H. Dyer and D. E. Edmunds, *From Real to Complex Analysis*,
Springer Undergraduate Mathematics Series, DOI: 10.1007/978-3-319-06209-9_2,

2.1 Basic Definitions

First we recall certain fundamental properties of real numbers: for all $x, y, z \in \mathbf{R}$,

(i) $|x-y| \geq 0$; $|x-y| = 0$ if, and only if, $x = y$;
(ii) $|x-y| = |y-x|$;
(iii) $|x-y| \leq |x-z| + |z-y|$.

The quantity $|x-y|$ is naturally thought of as the distance between the real numbers x and y. We seek to generalise all this, replacing $\mathbf{R}$ by an arbitrary non-empty set and $|x-y|$ by a function of x and y which satisfies axioms based on (i), (ii) and (iii). This is done, not simply as an exercise in the axiomatic approach, but because the structure obtained will enable us to solve many apparently different problems with the same technique.

Definition 2.1.1 Let X be a non-empty set and let $d : X \times X \to \mathbf{R}$ be such that for all $x, y, z \in X$,

(i) $d(x, y) \geq 0$; $d(x, y) = 0$ if, and only if, $x = y$;
(ii) $d(x, y) = d(y, x)$ (the symmetry property);
(iii) $d(x, y) \leq d(x, z) + d(z, y)$ (the triangle inequality).

The function d is called a **metric** or **distance function** on X; the pair (X, d) is called a **metric space**; when no ambiguity is possible we shall, for simplicity, often refer to X, rather than (X, d), as a metric space.

To illustrate this definition we give a variety of examples.

Example 2.1.2

(i) $X = \mathbf{R}$, $d(x, y) = |x-y|$: this was our prototype; d is called the **usual metric** on $\mathbf{R}$.
(ii) $X = \mathbf{R}$, $d(x, y) = |x-y| \,/\, (1 + |x-y|)$. To check that the triangle inequality holds, we observe that for all $x, y, z \in \mathbf{R}$,

$$d(x, y) = 1 - \frac{1}{1 + |x-y|} \leq 1 - \frac{1}{1 + |x-z| + |z-y|}$$

$$= \frac{|x-z| + |z-y|}{1 + |x-z| + |z-y|} \leq d(x, z) + d(z, y).$$

As the other properties required of a metric obviously hold, d is a metric.
(iii) Let $n \in \mathbf{N}$ and take

$$X = \mathbf{R}^n = \{x = (x_1, ..., x_n) = (x_i) : x_i \in \mathbf{R} \text{ for } i = 1, ..., n\}.$$

Various metrics can be defined on this set in a natural way: some of the most common are $d_p (1 \le p < \infty)$ and d_∞, where

$$d_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}, \quad 1 \le p < \infty,$$

$$d_\infty(x, y) = \max \{|x_i - y_i| : i = 1, ..., n\}.$$

The metric d_2 is usually referred to as the **Euclidean** metric on $\mathbf{R}^n$; when $n = 1$, all these metrics coincide. That each makes $\mathbf{R}^n$ into a metric space is clear, save perhaps for the proof of the triangle inequality for d_p, $1 \le p < \infty$. This follows from the Minkowski inequality (see Exercise 2.1.45/1), in view of which we see that for all $x, y, z \in \mathbf{R}^n$,

$$\begin{aligned} d_p (x, y) &= \left(\sum_{i=1}^{n} |(x_i - z_i) + (z_i - y_i)|^p \right)^{1/p} \\ &\le \left(\sum_{i=1}^{n} |x_i - z_i|^p \right)^{1/p} + \left(\sum_{i=1}^{n} |z_i - y_i|^p \right)^{1/p} \\ &= d_p(x, z) + d_p(z, y). \end{aligned}$$

We repeat that these examples illustrate the important fact that the same set may be endowed with different metrics. Note that for all $x, y \in \mathbf{R}^n$, $d_\infty (x, y) = \lim_{p\to\infty} d_p (x, y)$: this follows from the obvious inequalities

$$d_\infty (x, y) \le d_p (x, y) \le n^{1/p} d_\infty (x, y), \ 1 \le p < \infty.$$

(iv) Let X be any non-empty set and define $d : X \times X \to \mathbf{R}$ by the rule that

$$d(x, y) = \begin{cases} 1, & x \ne y, \\ 0, & x = y. \end{cases}$$

It is easy to check that d is a metric: (X, d) is called the **discrete metric space associated with** X, d being the **discrete metric** on X. This example is not only simple and a little surprising (going against our intuition about distances), but is also most useful as a source of counterexamples to rash conjectures about metric spaces.

In what follows, if $\mathbf{R}^n$ is referred to as a metric space without any metric being specified, then the Euclidean metric is to be assumed. When $n = 1$, identifying $\mathbf{R}^1$ and $\mathbf{R}$, this is the usual or standard metric.

(v) Let $a, b \in \mathbf{R}$, $a < b$, $I = [a, b]$, and let $X = C(I)$, the set of all continuous real-valued functions on I; for each $f, g \in C(I)$ define

$$\rho_1(f, g) = \int_a^b |f(t) - g(t)|\, dt,$$

$$\rho_\infty(f, g) = \max\{|f(t) - g(t)| : t \in [a, b]\}.$$

First note that since $|f - g|$ is a continuous, real-valued function on the closed, bounded interval I, both $\rho_1(f, g)$ and $\rho_\infty(f, g)$ are well-defined real numbers. It is now routine to check that both ρ_1 and ρ_∞ satisfy all the axioms (i), (ii) and (iii) of Definition 2.1.1: note in particular that in view of Theorem 1.3.2 (d), $\rho_1(f, g) = 0$ implies that $f = g$. Hence ρ_1 and ρ_∞ are metrics on $C(I)$. For details of a whole scale of metrics ρ_p $(1 \le p < \infty)$ on $C(I)$ see Exercise 2.1.45/2.

(vi) Next we give an example similar to $(\mathbf{R}^n, d_p)$ but in which the elements of the space are certain infinite sequences. That is, we let

$$X = \left\{ x = (x_i)_{i \in \mathbf{N}} : x_i \in \mathbf{R} \text{ for all } i \in \mathbf{N},\ \sum_1^\infty |x_i|^p < \infty \right\},\ 1 \le p < \infty,$$

and define d by

$$d(x, y) = \left(\sum_1^\infty |x_i - y_i|^p \right)^{1/p} \quad \text{for all } x, y \in X.$$

To show that (X, d) is a metric space, it is first necessary to verify that d is well-defined; that is, that $d(x, y) < \infty$ for all $x, y \in X$: in previous examples this has been rather obvious. For each $n \in \mathbf{N}$ we have, by Minkowski's inequality,

$$\left(\sum_1^n |x_i - y_i|^p \right)^{1/p} \le \left(\sum_1^n |x_i|^p \right)^{1/p} + \left(\sum_1^n |y_i|^p \right)^{1/p}$$

$$\le \left(\sum_1^\infty |x_i|^p \right)^{1/p} + \left(\sum_1^\infty |y_i|^p \right)^{1/p}.$$

Hence $d(x, y) < \infty$ and

$$\left(\sum_1^\infty |x_i - y_i|^p\right)^{1/p} \le \left(\sum_1^\infty |x_i|^p\right)^{1/p} + \left(\sum_1^\infty |y_i|^p\right)^{1/p}.$$

The triangle inequality now follows from this generalised version of Minkowski's inequality. To verify the remaining axioms is trivial.

This particular set X is usually referred to as ℓ_p.

(vii) Let $a, b \in \mathbf{R}$, $a < b$, $I = [a, b]$ and let $X = \mathscr{B}(I)$, the set of all bounded, real-valued functions on I, with d defined by

$$d(f, g) = \sup\{|f(t) - g(t)| : t \in I\} \text{ when } f, g \in \mathscr{B}(I).$$

It is easy to verify that d is a metric on $\mathscr{B}(I)$: axioms (i) and (ii) obviously hold, and if $f, g, h \in \mathscr{B}(I)$, then

$$\begin{aligned} d(f, g) &= \sup\{|f(t) - h(t) + h(t) - g(t)| : t \in I\} \\ &\le \sup\{|f(t) - h(t)| + |h(t) - g(t)| : t \in I\} \\ &\le \sup\{|f(t) - h(t)| : t \in I\} + \sup\{|h(t) - g(t)| : t \in I\} \\ &= d(f, h) + d(h, g), \end{aligned}$$

so that axiom (iii) also holds.

Since $C(I) \subset \mathscr{R}(I) \subset \mathscr{B}(I)$, we may regard $C(I)$ and $\mathscr{R}(I)$ as metric spaces, each with the metric inherited from $\mathscr{B}(I)$, that is, with the metrics $d\,|_{C(I)\times C(I)}$ and $d\,|_{\mathscr{R}(I)\times\mathscr{R}(I)}$ respectively.

(viii) Let (X, d) be a metric space and let Y be any non-empty subset of X; let d_Y be the restriction of d to $Y \times Y$. Then (Y, d_Y) is a metric space. Example 2.1.2 (vii) illustrates this most useful principle. In the case of any subset Y of $\mathbf{R}^n$, we shall for simplicity adopt the convention that if no metric is specified, Y is assumed to be endowed with the Euclidean metric inherited from $\mathbf{R}^n$.

(ix) Let (X_1, d_1), ..., (X_n, d_n) be metric spaces. The product space

$$X_1 \times ... \times X_n = \prod_{i=1}^n X_i = \{(x_1, ..., x_n) : x_i \in X_i \text{ for } i = 1, ..., n\}$$

may be made into a metric space by endowing it with the metric d, where

$$d(x, y) = \left\{\sum_{i=1}^n d_i^2(x_i, y_i)\right\}^{1/2} \text{ if } x = (x_i), y = (y_i).$$

This may be established just as it was shown in Example 2.1.2 (iii) that $(\mathbf{R}^n, d_2)$ is a metric space.

(x) Let X be a vector space over $\mathbf{R}$. Let $\|\cdot\| : X \to \mathbf{R}$ be a map such that

(a) $\|x\| = 0$ if, and only if, $x = 0$;

(b) $\|\alpha x\| = |\alpha| \, \|x\|$ for all $x \in X$ and all $\alpha \in \mathbf{R}$;

(c) $\|x + y\| \le \|x\| + \|y\|$ for all $x, y \in X$.

Such a map is said to be a **norm** on X. Any such norm generates a metric d on X given by

$$d(x, y) = \|x - y\| .$$

In several of the examples of metric spaces given above, namely (i), (iii), (v), (vi) and (vii), the underlying set X may be viewed as a real linear space and the metric is generated by a norm given by

$$\|x\| = d(x, 0).$$

We now introduce some particularly important subsets of a metric space.

Definition 2.1.3 Let (X, d) be a metric space. Given any $x \in X$ and any $r > 0$, let $B(x, r) = \{y \in X : d(x, y) < r\} : B(x, r)$ is called the **open ball in** X **with centre** x **and radius** r. A subset G of X is called **open** if given any $x \in G$, there exists $r > 0$ (depending upon x) such that $B(x, r) \subset G$.

Example 2.1.4

(i) Take $X = \mathbf{R}$ and let d be the usual metric given by $d(x, y) = |x - y|$ $(x, y \in \mathbf{R})$, so that $B(x, r) = (x - r, x + r)$. Then $(0, 1)$ is open, for given any $x \in (0, 1), B(x, \min\{x, 1 - x\}) \subset (0, 1)$; similarly, (a, b) is open for all $a \in \{-\infty\} \cup \mathbf{R}$ and all $b \in \mathbf{R} \cup \{+\infty\}$ with $a < b$. However, if $a, b \in \mathbf{R}$ and $a < b$, then $[a, b]$ is not open, for no matter what $r > 0$ we choose, $B(a, r)$ is not contained in $[a, b]$; similarly, $[a, b)$ and $(a, b]$ are not open.

(ii) In any metric space (X, d), X is plainly open; so is $\emptyset$, for since $\emptyset$ has no points, the statement 'for all $x \in \emptyset$, $B(x, r) \subset \emptyset$ for all $r > 0$' is true!

(iii) Let (X, d) be any metric space, let $x \in X$ and $r > 0$. Then $B(x, r)$ is open: this justifies our description of $B(x, r)$ as the *open* ball with centre x and radius r. To prove this, let $y \in B(x, r)$ and put $\varepsilon = r - d(x, y) > 0$. Then $B(y, \varepsilon) \subset B(x, r)$, for if $z \in B(y, \varepsilon)$, then

$$d(z, x) \le d(z, y) + d(y, x) < \varepsilon + d(x, y) = r.$$

(iv) Let $X = \mathbf{R}^2$ and let d be the metric d_2 of Example 2.1.2 (iii); that is,

$$d((x_1, x_2), (y_1, y_2)) = \left\{(x_1 - y_1)^2 + (x_2 - y_2)^2\right\}^{1/2} .$$

The set

$$S = \{(x_1, x_2) : x_2 > x_1\}$$

is open in $(\mathbf{R}^2, d)$, for given any $(x_1, x_2) \in S$, $B((x_1, x_2), (x_2 - x_1)/\sqrt{2}) \subset S$.

(v) Let $X = \mathbf{R}^2$ and consider the metrics d_1, d_2, d_∞ of Example 2.1.2 (iii) on $\mathbf{R}^2$. In $(\mathbf{R}^2, d_\infty)$, the open ball with centre (0, 0) and radius 1 is

$$\{(x_1, x_2) : \max\{|x_1|, |x_2|\} < 1\}.$$

In $\left(\mathbf{R}^2, d_1\right)$ the same open ball is $\{(x_1, x_2) : |x_1| + |x_2| < 1\}$, while in $\left(\mathbf{R}^2, d_2\right)$ this open ball has the more familiar specification $\left\{(x_1, x_2) : x_1^2 + x_2^2 < 1\right\}$. The reader is invited to sketch these three open balls.

(vi) In $\left(\mathbf{R}^2, d_2\right)$ the set $\mathbf{Q} \times \mathbf{Q} = \{(q_1, q_2) : q_1, q_2 \text{ rational}\}$ is not open, for given any $r > 0$, $\left(\sqrt{2}/n, 0\right) \in B\left((0, 0), r\right)$ for all sufficiently large $n \in \mathbf{N}$.

Some basic properties of open sets are given by the following Lemma.

Lemma 2.1.5 *Let (X, d) be a metric space.*

(i) *Every union of open subsets of X is open.*
(ii) *The intersection of every finite family of open subsets of X is open.*
(iii) *Let Y be a non-empty subset of X and let d_Y be the restriction of d to $Y \times Y$. Then U is an open subset of (Y, d_Y) if, and only if, there is an open subset V of (X, d) such that $U = V \cap Y$.*

Proof

(i) Let $\mathscr{U}$ be any family of open subsets of X and put $G = \bigcup \mathscr{U}$. If $G = \emptyset$ there is nothing to prove. Suppose $G \neq \emptyset$ and let $x \in G$. Then $x \in U$ for some $U \in \mathscr{U}$, and as U is open, there exists $r > 0$ such that $B(x, r) \subset U \subset G$. Hence G is open.
(ii) Let $\mathscr{U}$ be a finite family of open sets and put $F = \bigcap \mathscr{U}$. If $\mathscr{U} = \emptyset$, then $F = X$ and there is nothing to prove; again there is nothing to prove if $F = \emptyset$. Suppose $\mathscr{U} \neq \emptyset, F \neq \emptyset$ and let $x \in F$. Then $x \in U$ for all $U \in \mathscr{U}$; hence there exists $r_U > 0$ such that $B(x, r_U) \subset U$ for all $U \in \mathscr{U}$. Put $r = \min\{r_U : U \in \mathscr{U}\} : r > 0$ as $\mathscr{U}$ is a finite family, and so $B(x, r) \subset U$ for all $U \in \mathscr{U}$. Thus $B(x, r) \subset F$, and hence F is open.
(iii) If U is open in (Y, d_Y) then given any $u \in U$, there exists $r_u > 0$ such that $\{x \in Y : d(u, x) < r_u\} \subset U$; thus

$$U = \bigcup_{u \in U} \{x \in Y : d(u, x) < r_u\} = V \cap Y,$$

where

$$V = \bigcup_{u \in U} \{x \in X : d(u, x) < r_u\}$$

is open in (X, d).

Conversely, suppose $U = V \cap Y$, where V is open in (X, d). Then given any $u \in U$, there exists $r_u > 0$ such that

$$\{x \in Y : d(x, u) < r_u\} = Y \cap \{x \in X : d(x, u) < r_u\} \subset Y \cap V,$$

and so U is open in (Y, d_Y). □

Note that the intersection of infinitely many open sets need not be open. For if $X = \mathbf{R}$ and d is the usual metric on $\mathbf{R}$ (so that $d(x, y) = |x - y|$ for all $x, y \in \mathbf{R}$), then $\cap_{n=1}^{\infty} (-1/n, 1/n) = \{0\}$, which is not open in $(\mathbf{R}, d)$.

In general, not all subsets of a metric space are open: see Example 2.1.4 (i). We can, however, associate with each set a largest open subset: $(0, 1)$ is the largest open subset of $[0, 1)$ in $\mathbf{R}$, endowed with the usual metric, for instance.

Definition 2.1.6 Let (X, d) be a metric space and let $A \subset X$. The **interior** of A is defined to be the set

$$\overset{o}{A} = \bigcup \{G : G \subset A \textit{ and } G \textit{ is open in } X\}.$$

A point is said to be an **interior point** of A if it belongs to $\overset{o}{A}$.

Note that $\overset{o}{A}$ is the union of all the open sets contained in A; in view of Lemma 2.1.5 (i), it is plainly the largest open subset of A.

Example 2.1.7 Let $X = \mathbf{R}$ and let d be the usual metric on $\mathbf{R}$. The interior of $[0, 1] \cup \{67\}$ is $(0, 1)$; that of $\mathbf{N}$ is $\emptyset$.

Lemma 2.1.8 *A subset A of a metric space (X, d) is open if, and only if, $A = \overset{o}{A}$.*

Proof If A is open, then $A \subset \overset{o}{A} \subset A$, and so $A = \overset{o}{A}$. Conversely, if $A = \overset{o}{A}$, then since $\overset{o}{A}$ is open so is A. □

Dual to the notion of an open set is that of a closed set.

Definition 2.1.9 A subset A of a metric space X is **closed** if $X \backslash A$ is open.

Example 2.1.10

(i) In any metric space (X, d), both X and $\emptyset$ are closed (and open!). Moreover, given any $a \in X$, $\{a\}$ is closed, for given any $b \in X \backslash \{a\}$, $B\left(b, \frac{1}{2} d(a, b)\right) \subset X \backslash \{a\}$, so that $X \backslash \{a\}$ is open.

(ii) In $\mathbf{R}$, with the usual metric, $[a, b]$ is closed, for

$$\mathbf{R} \backslash [a, b] = (-\infty, a) \cup (b, \infty)$$

is open, as it is the union of two open sets.

(iii) The set $A = \{y \in X : d(x, y) \le r\}$ (x being a given point of X and r being a given positive number) is called the **closed ball with centre** x **and radius** r. This set is closed, for if $z \in X\backslash A$, then $B(z, d(z, x) - r) \subset X\backslash A$, which shows that $X\backslash A$ is open.

Lemma 2.1.11 *In any metric space, arbitrary intersections and finite unions of closed sets are closed.*

Proof Let $\Im$ be a collection of closed sets. Then by De Morgan's rules and Lemma 2.1.5,

$$^c\big(\bigcap_{F\in\Im} F\big) = \bigcup_{F\in\Im} {}^cF$$

is open; hence $\bigcap_{F\in\Im} F$ is closed. Let $F_1, \ldots, F_n$ be closed sets. Then

$$^c(F_1 \cup \cdots \cup F_n) = {}^cF_1 \cap \cdots \cap {}^cF_n,$$

a finite intersection of open sets. Thus $^c(F_1 \cup \cdots \cup F_n)$ is open, by Lemma 2.1.5; hence $(F_1 \cup \cdots \cup F_n)$ is closed. □

Dual to the notion of the interior is that of the closure of a set.

Definition 2.1.12 The **closure** $\overline{A}$ of a subset A of a metric space X is the intersection of all closed sets in X which contain A.

In view of Lemma 2.1.11, $\overline{A}$ is the smallest closed set which contains A. Two simple, but useful, lemmas now follow.

Lemma 2.1.13 *Let A be a subset of a metric space. Then A is closed if, and only if, $A = \overline{A}$. Moreover, $\overline{\overline{A}} = \overline{A}$.*

Proof If $A = \overline{A}$, then since $\overline{A}$ is closed, so is A. Conversely, if A is closed then it is the smallest closed set which contains A, and hence $A = \overline{A}$. Since $\overline{A}$ is closed, it now follows that $\overline{\overline{A}} = \overline{A}$. □

Lemma 2.1.14 *Let A be a subset of a metric space X. Then*

$$^c(\overset{o}{A}) = \overline{{}^cA} \quad \textit{and} \quad {}^c(\overline{A}) = \overset{o}{\widehat{{}^cA}}\,.$$

Proof A point x belongs to $^c(\overset{o}{A})$ if, and only if, x fails to belong to any open set $G \subset A$; and this is so if, and only if, $x \in F$ for all closed $F \supset {}^c A$, which is equivalent to the statement that $x \in \overline{{}^cA}$.

The second identity follows from the first on replacing A by ^cA. Alternatively, note that

$$^c(\overline{A}) = {}^c\left(\bigcap\nolimits_{F\supset A, F \text{ closed}} F\right) = \bigcup\nolimits_{F\supset A, F\text{ closed}} {}^cF = \bigcup\nolimits_{O \text{ open}, O\subset {}^cA} O = \overset{o}{\widehat{{}^cA}}. \qquad \square$$

For economy of expression we need the following definition.

Definition 2.1.15 Let X be a metric space and let $a \in X$. Any open set containing a will be called a **neighbourhood** of a.

A simple example of a neighbourhood of a is given by the open ball $B(a, r)$ centred at a and with radius r; every neighbourhood of a contains such a ball. In terms of neighbourhoods we can give a useful characterisation of the points of the closure of a set.

Lemma 2.1.16 *Let A be a subset of a metric space X. Then $x \in \overline{A}$ if, and only if, every neighbourhood V of x has non-empty intersection with A.*

Proof The statement that for all neighbourhoods V of x we have $V \cap A \neq \emptyset$ is equivalent to saying that $x \notin \overset{o}{\widehat{{}^cA}} = {}^c(\overline{A})$, by Lemma 2.1.14. □

Definition 2.1.17 Let A be a subset of a metric space X. The **boundary** ∂A of A is defined to be $\overline{A} \backslash \overset{o}{A}$.

Note that by Lemma 2.1.14,

$$\partial A = \overline{A} \cap \overline{{}^cA} = \overline{{}^cA} \backslash \overset{o}{\widehat{{}^cA}}\ .$$

Example 2.1.18

(i) Let $X = \mathbf{R}$, endowed with the usual metric. The boundary of $[0, 1]$ is $\{0, 1\}$, that of $\mathbf{Q}$ and $\mathbf{R}\backslash\mathbf{Q}$ is $\mathbf{R}$.
(ii) Let X be a discrete metric space and let $x \in X$. Then $\partial B(x, r) = \emptyset$ for all $r > 0$. This contrasts sharply with the situation in $\mathbf{R}^n$, equipped with the Euclidean metric: in this setting $\partial B(x, r) = \left\{ y \in \mathbf{R}^n : \sum_{i=1}^{n} (x_i - y_i)^2 = r^2 \right\}$.

Now that we have introduced the basic ideas concerning subsets of a metric space we turn to the convergence of sequences.

Definition 2.1.19 A sequence (x_n) in a metric space (X, d) is said to **converge to a point** $x \in X$ if, and only if, given any $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that $d(x, x_n) < \varepsilon$ if $n \geq N$; we write this as $x_n \to x$, $\lim_{n\to\infty} x_n = x$ or $d(x, x_n) \to 0$ as $n \to \infty$. A sequence (x_n) in X is said to be **convergent** if, and only if, there exists $x \in X$ such that $x_n \to x$; we also say in this case that (x_n) has **limit** x.

Note that $x_n \to x$ if, and only if, given any neighbourhood V of x, there exists $N \in \mathbf{N}$ such that $x_n \in V$ for all $n \geq N$.

Lemma 2.1.20 *Let (x_n) be a sequence in a metric space (X, d). Then (x_n) converges to at most one point.*

Proof Suppose that $x_n \to x$ and $x_n \to y$. Then

$$d(x, y) \leq d(x, x_n) + d(x_n, y) \to 0 \text{ as } n \to \infty.$$

Hence $d(x, y) = 0$, and so $x = y$. □

Of course, a sequence may well not converge to any point.

Example 2.1.21

(i) Let $X = \mathbf{R}^n$ and let d be the Euclidean metric on $\mathbf{R}^n$ (see Example 2.1.2 (iii)); let $(x^{(m)})_{m\in\mathbf{N}}$ be a sequence in $\mathbf{R}^n$, with $x^{(m)} = (x_1^{(m)}, \ldots, x_n^{(m)})$. The sequence $(x^{(m)})$ converges to $x = (x_i)$ in $\mathbf{R}^n$ if, and only if,

$$\sum_{i=1}^{n}(x_i - x_i^{(m)})^2 \to 0 \text{ as } m \to \infty;$$

from this it is clear that $(x^{(m)})$ converges to x in $\mathbf{R}^n$ if, and only if, $(x_i^{(m)})$ converges to x_i as $m \to \infty$, for each $i \in \{1, \ldots, n\}$.

(ii) Let (X, d) be a discrete metric space. Then a sequence (x_n) in X is convergent if, and only if, it is eventually constant; that is if, and only if, there exists $N \in \mathbf{N}$ such that $x_n = x_N$ for all $n \geq N$. For if (x_n) is convergent in X, there exists $x \in X$ and $N \in \mathbf{N}$ such that $d(x, x_n) < 1$ for all $n \geq N$, so that $x_n = x$ for all $n \geq N$. The converse is obvious.

Convergent sequences of real numbers are bounded. Given an appropriate extension of the definition of boundedness, the same is true in a general metric space.

Definition 2.1.22 Let (X, d) be a metric space. A non-empty set $A \subset X$ is said to be **bounded** if there is a real number M such that

$$d(x, y) \leq M \; (x, y \in A);$$

otherwise, A is said to be **unbounded**. The extended real number

$$\operatorname{diam}(A) := \sup\{d(x, y) : x, y \in A\}$$

is called the **diameter** of A.

Note that the set A is bounded if, and only if, $\operatorname{diam}(A) < \infty$.

Lemma 2.1.23 *Let (x_n) be a convergent sequence in a metric space (X, d). Then $\{x_n : n \in \mathbf{N}\}$ is bounded.*

Proof Suppose that $\lim_{n\to\infty} x_n = x$. Then there exists $N \in \mathbf{N}$ such that for all $n \geq N$, $d(x, x_n) < 1$. Put $r = \max\{1, d(x, x_1), \ldots, d(x, x_{N-1})\}$. Then $d(x, x_n) \leq r$ for all $n \in \mathbf{N}$; further,

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) \leq 2r \quad (m, n \in \mathbf{N}).$$

Thus $\{x_n : n \in \mathbf{N}\}$ is bounded. □

We can now give a most useful characterisation of the closure of a subset of a metric space.

Lemma 2.1.24 *Let A be any subset of a metric space X. Then $x \in \overline{A}$ if, and only if, there is a sequence (x_n) of points of A such that* $\lim_{n\to\infty} x_n = x$.

Proof Suppose there is a sequence (x_n) in A such that $x_n \to x$ as $n \to \infty$. Then for all $r > 0$, $A \cap B(x, r) \neq \emptyset$; and so by Lemma 2.1.16, $x \in \overline{A}$.

Conversely, suppose that $x \in \overline{A}$. Then by Lemma 2.1.16, for all $n \in \mathbf{N}$ we have $B(x, \frac{1}{n}) \cap A \neq \emptyset$. Appeal to the countable axiom of choice (Axiom A.5.2) gives the existence of a sequence (x_n) with $x_n \in B(x, \frac{1}{n}) \cap A$ for all $n \in \mathbf{N}$; plainly $x_n \to x$. □

To conclude this rapid discussion of sequences we introduce the notion of a point of accumulation of a set.

Definition 2.1.25 Let A be a subset of a metric space X. A point $x \in X$ is called an **accumulation point of** A, or a **limit point of** A, if given any neighbourhood V of x, there exists $a \in A \cap V$ with $a \neq x$.

Note the difference between a point of accumulation of A and a point in $\overline{A}$: every point of accumulation of A is evidently in $\overline{A}$, but the converse is false. For example, with $X = \mathbf{R}$ endowed with the usual metric and $A = (0, 1) \cup \{2\}$, the point 2 belongs to $\overline{A}$ but is not a point of accumulation of A as $B(2, 1)$ contains no point of A distinct from 2.

Lemma 2.1.26 *Let $A \subset X$. Then x is a point of accumulation of A if, and only if, there is a sequence (x_n) of distinct points of A with $x_n \to x$ as $n \to \infty$.*

Proof Let x be a point of accumulation of A. Then given any $n \in \mathbf{N}$, there exists $x_n \in A \cap B(x, \frac{1}{n})$, $x_n \neq x$; this gives a sequence (x_n) of points of A which converges to x, with each $x_n \neq x$. The difficulty is that the points of this sequence may not be distinct, and to overcome this we proceed as follows, noting that there must be infinitely many distinct points in the sequence, for otherwise the sequence could not converge to x. Define $m : \mathbf{N} \to \mathbf{N}$ by $m(1) = 1$, $m(k + 1)$ = least integer p such that $x_p \notin \left\{x_{m(1)}, x_{m(2)}, \ldots, x_{m(k)}\right\}$ $(k \geq 1)$; thus $m(2)$ = least p such that $x_p \neq x_1$. Then $(x_{m(n)})_{n\in\mathbf{N}}$ is a subsequence of (x_n) consisting of distinct points of A, and $\lim_{n\to\infty} x_{m(n)} = x$. The converse is obvious. □

2.1.1 Continuous Functions

Definition 2.1.27 Let (X_1, d_1), (X_2, d_2) be metric spaces. A map $f : X_1 \to X_2$ is said to be **continuous at** $x \in X_1$ if given any $\varepsilon > 0$, there exists $\delta > 0$ such that $d_2(f(y), f(x)) < \varepsilon$ if $d_1(x, y) < \delta$. (In general, δ depends upon x and ε.) If f is continuous at each point of X_1, it is said to be **continuous (on** X_1). If given any $\varepsilon > 0$, there exists $\delta > 0$ (depending only on ε) such that $d_2(f(y), f(x)) < \varepsilon$ whenever $d_1(x, y) < \delta$, then f is called **uniformly continuous on** X_1.

This definition is the obvious extension of the ε, δ definition of continuity and uniform continuity for maps from subsets of **R** to **R** given in Chap. 1. However, in the wider context of metric spaces it is desirable to have other characterisations of continuity, and we now deal with this, beginning with the local property (that is, continuity at a point) and then turning to the global position (continuity on the whole space).

Lemma 2.1.28 *Let (X_1, d_1) and (X_2, d_2) be metric spaces, let $f : X_1 \to X_2$ and let $x \in X_1$. Then the following three statements are equivalent:*

(i) *f is continuous at x;*
(ii) *given any neighbourhood V of $f(x)$, there is a neighbourhood U of x such that $f(U) \subset V$;*
(iii) $\lim_{n\to\infty} f(x_n) = f(x)$ *if $x_n \to x$.*

Proof To prove that (i) implies (ii), let V be a neighbourhood of $f(x)$ and let $\varepsilon > 0$ be such that $B(f(x), \varepsilon) \subset V$. Since f is continuous at x, there exists $\delta > 0$ such that $f(y) \in B(f(x), \varepsilon)$ if $y \in B(x, \delta)$; thus $f(B(x, \delta)) \subset B(f(x), \varepsilon)$ and (ii) holds with $U = B(x, \delta)$. Next we show that (ii) implies (iii). Suppose that $x_n \to x$ in X_1 and let V be a neighbourhood of $f(x)$. As (ii) holds, there is a neighbourhood U of x such that $f(U) \subset V$; and there exists $N \in \mathbf{N}$ such that $x_n \in U$ if $n \geq N$. Hence $f(x_n) \in V$ for all $n \geq N$, which means that $f(x_n) \to f(x)$ as $n \to \infty$.

Finally, to prove that (iii) implies (i), suppose that (iii) holds but (i) is not true. Then there is an $\varepsilon > 0$ such that given any $n \in \mathbf{N}$, there exists $x_n \in X_1$ such that $d_1(x, x_n) < 1/n$ while $d_2(f(x), f(x_n)) \geq \varepsilon$; and so $x_n \to x$ but $f(x_n) \nrightarrow f(x)$, which contradicts (iii). □

Lemma 2.1.29 *Let X_1, X_2 and X_3 be metric spaces, let $x \in X_1$, let $f \colon X_1 \to X_2$ be continuous at x and let $g : X_2 \to X_3$ be continuous at $f(x)$. Then $h := g \circ f$ is continuous at x. If f and g are continuous on X_1 and X_2 respectively, then h is continuous on X_1.*

Proof Suppose $x_n \to x$ as $n \to \infty$. Then as f is continuous at x, $f(x_n) \to f(x)$; and as g is continuous at $f(x)$, $g(f(x_n)) \to g(f(x))$. By Lemma 2.1.28, h is continuous at x. The rest is obvious. □

Example 2.1.30

(i) Let $f : \mathbf{R}^m \to \mathbf{R}^n$, and suppose that $\mathbf{R}^m$ and $\mathbf{R}^n$ are endowed with the appropriate Euclidean metric. For each $x \in \mathbf{R}^m$ write $f(x) = (f_1(x), \ldots, f_n(x))$; we thus have defined functions $f_i : \mathbf{R}^m \to \mathbf{R}$ $(i = 1, \ldots, n)$, called the **coordinate functions of** f. It is now clear that f is continuous at $x_0 \in \mathbf{R}^m$ if, and only if, each f_i is continuous at x_0.
(ii) Just as in the case of maps from **R** to **R** it follows that if X is a metric space then sums and products of continuous maps from X to **R** are continuous; that is, if $x_0 \in X$ and $f_1, f_2 : X \to \mathbf{R}$ are continuous at x_0, then the maps $f_1 + f_2$ and $f_1 f_2$ (defined by $x \longmapsto f_1(x) + f_2(x)$ and $x \longmapsto f_1(x) f_2(x)$ respectively)

are continuous at x_0. Similarly, the map λf_1 (defined by $x \longmapsto \lambda f_1(x)$) is continuous at x_0, for all $\lambda \in \mathbf{R}$; and if $f_2(x_0) \neq 0$, then the map f_1/f_2 (defined by $x \longmapsto f_1(x)/f_2(x)$) is continuous at x_0. The proofs of all these assertions are identical to those of the corresponding assertions when $X = \mathbf{R}$ and which are familiar in elementary analysis.

It follows that every polynomial p on $\mathbf{R}^2$, where

$$p(x, y) = \sum_{m,n=0}^{N} a_{mn}x^m y^n (a_{mn} \in \mathbf{R}),$$

is continuous on $\mathbf{R}^2$; and any rational function f on $\mathbf{R}^2$, where $f(x, y) = p(x, y)/q(x, y)$ and p, q are polynomials with q never zero, is also continuous on $\mathbf{R}^2$.

(iii) Let $f : \mathbf{R}^2 \to \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} \frac{x^2-y^2}{x^2+y^2} & \text{if } (x, y) \neq (0, 0), \\ 0 & \text{if } (x, y) = (0, 0). \end{cases}$$

Reasoning as in (ii), f is continuous on $\mathbf{R}^2 \setminus \{0, 0\}$; it is not continuous at $(0, 0)$, for if $x \neq 0, f(x, 0) = 1 \nrightarrow 0 = f(0, 0)$ as $x \to 0$.

Functions between metric spaces commonly have points of discontinuity. As a tool for the investigation of discontinuity we introduce the concept of the oscillation of a function at a point.

Definition 2.1.31 Let X_1 and X_2 be metric spaces and let f be a map from X_1 to X_2. For each $x \in X_1$, let $\omega(x)$ be the extended real number defined by

$$\omega(x) = \inf\{\operatorname{diam}(f(U)) : U \textit{is a neighbourhood of } x\};$$

$\omega(x)$ is called the **oscillation of** f **at** x. The corresponding function ω is called the **oscillation function for** f.

Lemma 2.1.32 *Let X_1 and X_2 be metric spaces, f be a map from X_1 to X_2, and ω be the oscillation function for f. Then*

(a) *f is continuous at $x \in X_1$ if, and only if, $\omega(x) = 0$;*
(b) *for each real number α, the set $\{x \in X_1 : \omega(x) < \alpha\}$ is open in X_1.*

Proof (a) Let f be continuous at x and $\varepsilon > 0$. Then there exists $\delta > 0$ such that

$$f(B(x, \delta)) \subset B(f(x), \varepsilon).$$

Hence

$$\omega(x) \le \operatorname{diam}(f(B(x,\delta))) \le 2\varepsilon,$$

and it follows that $\omega(x) = 0$.

Conversely, let $\omega(x) = 0$ and $\varepsilon > 0$. There is a neighbourhood U of x such that $\operatorname{diam}(f(U)) < \varepsilon$; further, there exists $\delta > 0$ such that $B(x,\delta) \subset U$. Hence $f(B(x,\delta)) \subset B(f(x),\varepsilon)$ and f is continuous at x.
(b) Suppose $\alpha > 0$; the result is obvious otherwise. Let

$$E = \{x \in X_1 : \omega(x) < \alpha\}$$

and $y \in E$. Then there is a neighbourhood U of y such that $\operatorname{diam}(f(U)) < \alpha$. Now U is a neighbourhood of each of its points and thus $U \subset E$. It follows that $y \in \overset{o}{E}$, that $E \subset \overset{o}{E}$ and so E is open. □

Lemma 2.1.33 *Let X_1 and X_2 be metric spaces and let $f : X_1 \to X_2$. The following three statements are equivalent:*

(i) *f is continuous (on X_1);*
(ii) *if V is an open subset of X_2, $f^{-1}(V)$ is open in X_1;*
(iii) *if F is a closed subset of X_2, $f^{-1}(F)$ is closed in X_1.*

Proof To prove that (i) implies (ii), assume that (i) holds, let V be open in X_2 and let $x \in f^{-1}(V)$. As f is continuous at x, there is a neighbourhood $U(x)$ of x such that $f(U(x)) \subset V$; that is, $U(x) \subset f^{-1}(V)$. Thus $f^{-1}(V)$ contains a neighbourhood of each of its points and hence is open.

Next suppose that (ii) holds, let $x \in X_1$ and let V be a neighbourhood of $f(x)$. Then by (ii), $f^{-1}(V)$ is open; and $x \in f^{-1}(V)$. Thus $f^{-1}(V)$ is a neighbourhood of x and $f(f^{-1}(V)) \subset V$, which by Lemma 2.1.28 means that f is continuous at x. Since x is an arbitrary point of X_1, f must be continuous on X_1. Hence (i) and (ii) are equivalent.

Finally, (ii) and (iii) are equivalent, in view of the identity $X_1 \backslash f^{-1}(F) = f^{-1}(X_2 \backslash F)$ for all $F \subset X_2$. □

Remark 2.1.34

(i) In view of Lemma 2.1.33 it is easy to see that $f : X_1 \to X_2$ is continuous if, and only if, $f^{-1}(B)$ is open for all open balls $B \subset X_2$.
(ii) Suppose that f: $X_1 \to X_2$ is continuous and that U is an open subset of X_1. It does not follow that $f(U)$ is open in X_2. To illustrate this important point, let $X_1 = X_2 = \mathbf{R}$, endowed with the usual metric, define f: $\mathbf{R} \to \mathbf{R}$ by $f(x) = (1+x^2)^{-1}$ $(x \in \mathbf{R})$ and let $U = (-1,1)$. Then f is continuous, U is open but $f(U) = \left(\frac{1}{2}, 1\right]$, which is not open. Similarly, it does not follow that the image of a closed set under a continuous map is closed.

Lemma 2.1.33 enables us to prove a simple and most useful result, often called the glueing lemma because it shows that under appropriate conditions, two continuous

functions defined on subsets of a metric space may be 'glued together' to form a continuous function on the union of those subsets. Frequent reference to this lemma will be made in Sect. 2.5 and in our treatment of the Jordan curve theorem in Sect. 3.9.

Lemma 2.1.35 *Let X and Y be metric spaces and suppose that $X = A \cup B$, where A and B are non-empty and either both open or both closed. Let $f : A \to Y$ and $g : B \to Y$ be continuous (A and B are assumed equipped with the metric inherited from X), suppose that $f(x) = g(x)$ for all $x \in A \cap B$, and define $h : X \to Y$ by*

$$h(x) = \begin{cases} f(x), & x \in A, \\ g(x), & x \in B. \end{cases}$$

Then h is continuous.

Proof (i) Suppose A and B are both open in X. Let $\mathscr{O}$ be an open set in Y. By Lemma 2.1.33, $f^{-1}(\mathscr{O})$ and $g^{-1}(\mathscr{O})$ are open in A and B, respectively. Thus, by Lemma 2.1.5, there exist sets U, V open in X such that

$$f^{-1}(\mathscr{O}) = U \cap A, \ g^{-1}(\mathscr{O}) = V \cap B;$$

the sets $f^{-1}(\mathscr{O})$ and $g^{-1}(\mathscr{O})$ are open in X; and since $h^{-1}(\mathscr{O}) = f^{-1}(\mathscr{O}) \cup g^{-1}(\mathscr{O})$, $h^{-1}(\mathscr{O})$ is open in X. The continuity of h follows by further appeal to Lemma 2.1.33.

(ii) Suppose A and B are both closed in X and let F be a closed set in Y. By Lemma 2.1.33, $f^{-1}(F)$ and $g^{-1}(F)$ are closed in A and B, respectively. By Lemma 2.1.5, since $A \backslash f^{-1}(F)$ is open in A, there exists a set U open in X such that $A \backslash f^{-1}(F) = A \cap U$; also,

$$X \backslash f^{-1}(F) = (X \backslash A) \cup (A \backslash f^{-1}(F)) = (X \backslash A) \cup U,$$

a set open in X. Thus $f^{-1}(F)$ is closed in X as, by similar reasoning, is $g^{-1}(F)$. Now $h^{-1}(F) = f^{-1}(F) \cup g^{-1}(F)$ and so, by Lemma 2.1.11, $h^{-1}(F)$ is closed in X. Finally, by Lemma 2.1.33, the continuity of h is proved.

Note that use of Exercise 2.1.45/17 yields a simpler proof of (ii), one identical in form with that given in case (i). □

2.1.2 Homeomorphisms

We now introduce the idea of homeomorphism, which enables a sensible classification of spaces to be made.

Definition 2.1.36 Let X_1, X_2 be metric spaces. A map $f : X_1 \to X_2$ is said to be a **homeomorphism** if it is a continuous bijection and f^{-1} is continuous. If such a map exists, X_1 and X_2 are said to be **homeomorphic**.

Remark 2.1.37

(i) Not every bijective continuous map is a homeomorphism; that is, the condition that f^{-1} be continuous is an essential part of the definition. For take $X_1 = \mathbf{R}$ endowed with the discrete metric, let $X_2 = \mathbf{R}$ endowed with the usual metric and let f be the identity map from X_1 to X_2; that is $f(x) = x$ for all $x \in \mathbf{R}$. Then f is not a homeomorphism: it is continuous, as $f^{-1}(O)$ is open in X_1 whenever O is open in X_2 (every subset of X_1 is open!); but f^{-1} is not continuous, for $(f^{-1})^{-1}([0, 1)) = [0, 1)$ is open in X_1 but not in X_2.

(ii) Let $f : X_1 \to X_2$ be a bijection. Then if f is a homeomorphism, a subset U of X_1 is open if, and only if, $f(U)$ is open in X_2: note that $U = (f^{-1})^{-1}(U)$. It follows that the open sets in a metric space may be put in one-to-one correspondence with the open sets in any metric space homeomorphic to it.

(iii) In general, homeomorphisms do not preserve distances. Thus let $X_1 = (0, 1)$, $X_2 = (1, \infty)$ and endow each set with the usual metric inherited from $\mathbf{R}$; let $f : X_1 \to X_2$ be defined by $f(x) = x^{-1}$ $(x \in X_1)$. Then f is plainly a homeomorphism, but if $x, y \in X_1$ and $x \neq y$, then $\left|x^{-1} - y^{-1}\right| \neq |x - y|$; that is, the distance between $f(x)$ and $f(y)$ differs from that between x and y. Homeomorphisms which do preserve distances are called **isometries**. We formalise this in the following definition.

Definition 2.1.38 Let (X_1, d_1) and (X_2, d_2) be metric spaces. A map $f : X_1 \to X_2$ is said to be an **isometry** if it is bijective and for all $x, y \in X_1$,

$$d_2(f(x), f(y)) = d_1(x, y).$$

If such a map exists, X_1 and X_2 are said to be **isometric**.

Example 2.1.39

(i) Consider $\mathbf{R}^n$, with the Euclidean metric, and the unit ball $B(0, 1)$ in $\mathbf{R}^n$, given the metric inherited from $\mathbf{R}^n$. Then $\mathbf{R}^n$ and $B(0, 1)$ are homeomorphic. To see this, let $f : \mathbf{R}^n \to B(0, 1)$ be defined by $f(x) = x/(1+|x|)$, where $|x| = \left(\sum_{i=1}^{n} x_i^2\right)^{\frac{1}{2}}$. Since $|f(x)| = |x|/(1+|x|) < 1$ for all $x \in \mathbf{R}^n$, it follows that $f(\mathbf{R}^n) \subset B(0, 1)$. Moreover, given any $y \in B(0, 1)$, the point $x := y/(1 - |y|)$ is the unique point mapped by f to y : thus f is a bijection and $f^{-1}(y) = y/(1 - |y|)$. It is clear that f and f^{-1} are continuous, and hence f is a homeomorphism.

(ii) For any $n \in \mathbf{N}$, let S^n be the unit sphere in $\mathbf{R}^{n+1}$ (endowed with the Euclidean metric); that is,

$$S^n = \left\{ (x_i) \in \mathbf{R}^{n+1} : \sum_{i=1}^{n+1} x_i^2 = 1 \right\}.$$

Then $S^2 \setminus \{(0, 0, 1)\}$ is homeomorphic to $\mathbf{R}^2$, each set being given the Euclidean metric. This follows by consideration of the stereographic projection P, where

$$P(x_1, x_2, x_3) = \left(\frac{x_1}{1 - x_3}, \frac{x_2}{1 - x_3} \right).$$

First note that P is obviously a continuous map of $S^2 \setminus \{(0, 0, 1)\}$ to $\mathbf{R}^2$. Moreover, P is bijective, for given any $(y_1, y_2) \in \mathbf{R}^2$, the equation $Px = (y_1, y_2)$ has the unique solution $x_1 = 2y_1/(y_1^2 + y_2^2 + 1)$, $x_2 = 2y_2/(y_1^2 + y_2^2 + 1)$, $x_3 = (y_1^2 + y_2^2 - 1)/(y_1^2 + y_2^2 + 1)$, since any possible solution $x = (x_1, x_2, x_3)$ must satisfy

$$(x_1^2 + x_2^2)/(1 - x_3)^2 = y_1^2 + y_2^2, \ (1 - x_3^2)/(1 - x_3)^2 = y_1^2 + y_2^2,$$

so that $(1 + x_3)/(1 - x_3) = y_1^2 + y_2^2$ and hence

$$x_3 = (y_1^2 + y_2^2 - 1)/(y_1^2 + y_2^2 + 1), \ x_1 = y_1(1 - x_3) = 2y_1/(y_1^2 + y_2^2 + 1),$$

$$x_2 = y_2(1 - x_3) = 2y_2/(y_1^2 + y_2^2 + 1).$$

The continuity of P^{-1} is now clear, and so P is a homeomorphism.

(iii) Let $S, Q \subset \mathbf{R}^2$ be a circle and a square, respectively, each given the Euclidean metric inherited from $\mathbf{R}^2$. Then S and Q are homeomorphic. To prove this it is enough to consider the case in which $S = S^1$ (see (ii) above) and Q is the square with centre O, of side 2 and with sides parallel to the coordinate axes. Define $\phi : Q \to S$ by $\phi(x, y) = (x, y)/\sqrt{(x^2 + y^2)}$; ϕ is plainly continuous. It is bijective, for given $(u, v) \in S$, there is a unique (x, y) in Q such that $\phi(x, y) = (u, v)$. In fact, $x/\sqrt{(x^2 + y^2)} = u, y/\sqrt{(x^2 + y^2)} = v$, so that if we put $x = r\cos\theta$, $y = r\sin\theta$, then $u = \cos\theta$, $v = \sin\theta$; moreover, $(x, y) \in Q$ if, and only if, $\max\{|x|, |y|\} = 1$, and so $\max\{r|u|, r|v|\} = 1$, which gives $r = 1/\max\{|u|, |v|\}$. Hence

$$x = u/\max\{|u|, |v|\}, \ y = v/\max\{|u|, |v|\},$$

which shows that $\phi^{-1}(u, v) = (u, v)/\max\{|u|, |v|\}$. Since ϕ^{-1} is plainly continuous, it follows that ϕ is a homeomorphism.

(iv) Let $\mathbf{R}^n$ be given the Euclidean metric. Then a map $g : \mathbf{R}^n \to \mathbf{R}^n$ is an isometry if, and only if, it is of the form

$$g(t) = x_0 + f(t) \ (t \in \mathbf{R}^n)$$

where $x_0 \in \mathbf{R}^n$ and $f : \mathbf{R}^n \to \mathbf{R}^n$ is linear, and orthogonal in the sense that for all $s, t \in \mathbf{R}^n$, $\langle f(s), f(t) \rangle = \langle s, t \rangle$, where $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$.

To prove this, first suppose that g is an isometry and let d be the Euclidean metric on $\mathbf{R}^n$. Put $f(t) = g(t) - g(0)$ $(t \in \mathbf{R}^n)$; then $f(0) = 0$, $d(f(s),$

$f(t)) = d(g(s), g(t)) = d(s, t)$ and $d(f(t), 0) = d(g(t), g(0)) = d(t, 0)$. For all $s, t \in \mathbf{R}^n$, it follows that since $d^2(f(s), f(t)) = d^2(s, t)$ we have

$$d^2(f(t), 0) - 2\langle f(t), f(s)\rangle + d^2(f(s), 0) = d^2(s, 0) - 2\langle s, t\rangle + d^2(t, 0),$$

and hence $\langle f(t), f(s)\rangle = \langle s, t\rangle$.

Thus f is orthogonal. To show that f is linear, let $s, t \in \mathbf{R}^n$ and $\alpha, \beta \in \mathbf{R}$. Then $d^2(f(\alpha s + \beta t), \alpha f(s) + \beta f(t))$ is given by

$$\begin{aligned}
&\langle f(\alpha s + \beta t) - \alpha f(s) - \beta f(t), f(\alpha s + \beta t) - \alpha f(s) - \beta f(t)\rangle \\
&= \langle f(\alpha s + \beta t), f(\alpha s + \beta t)\rangle + \alpha^2 \langle f(s), f(s)\rangle + \beta^2 \langle f(t), f(t)\rangle \\
&\quad + 2\alpha\beta \langle f(s), f(t)\rangle - 2\alpha \langle f(\alpha s + \beta t), f(s)\rangle - 2\beta \langle f(\alpha s + \beta t), f(t)\rangle \\
&= \langle \alpha s + \beta t, \alpha s + \beta t\rangle + \alpha^2 \langle s, s\rangle + \beta^2 \langle t, t\rangle + 2\alpha\beta \langle s, t\rangle \\
&\quad - 2\alpha \langle \alpha s + \beta t, s\rangle - 2\beta \langle \alpha s + \beta t, t\rangle \\
&= 2\langle \alpha s + \beta t, \alpha s + \beta t\rangle - 2\langle \alpha s + \beta t, \alpha s + \beta t\rangle = 0.
\end{aligned}$$

Hence f is linear.

Conversely, suppose that $g = x_0 + f$, where f is linear and orthogonal. Then for all $s, t \in \mathbf{R}^n$,

$$\begin{aligned}
d^2(g(s), g(t)) &= d^2(f(s), f(t)) = \langle f(s) - f(t), f(s) - f(t)\rangle \\
&= \langle f(s - t), f(s - t)\rangle = \langle s - t, s - t\rangle = d^2(s, t),
\end{aligned}$$

which shows that g is an isometry.

2.1.3 An Extension Theorem

Let A be a subspace of a metric space X and let $f : A \to \mathbf{R}$ be continuous. A natural question to ask is whether or not f has a continuous real-valued extension defined on all of X. That is to say, does there exist a continuous map $g : X \to \mathbf{R}$ such that for all $x \in A$, $g(x) = f(x)$? In general, the answer is negative: the map $x \longmapsto 1/x : (0, 1) \to \mathbf{R}$ is continuous, but it has no continuous extension even

to [0, 1] because such an extension would have to be bounded. However, for an affirmative answer it turns out that a sufficient condition on A is that it is a closed subspace of X. We establish this below, beginning with a few preliminaries.

Lemma 2.1.40 *Let A and B be non-empty subsets of a metric space (X, d). For each $x \in X$, let the **distance from** x **to** A, denoted by $d(x, A)$ or by $\operatorname{dist}(x, A)$ when the metric d is understood, be defined by*

$$d(x, A) = \inf \{d(x, a) : a \in A\} \quad (x \in X),$$

*and the **distance from** A **to** B be*

$$d(A, B) = \inf \{d(a, b) : a \in A, b \in B\}.$$

Then

(i) $d(A, B) = \inf \{d(a, B) : a \in A\}$;
(ii) $d(x, A) = 0$ *if, and only if,* $x \in \overline{A}$;
(iii) *the map* $x \longmapsto d(x, A) : X \to \mathbf{R}$ *is continuous; in fact, for all* $x, y \in X$,

$$|d(x, A) - d(y, A)| \leq d(x, y).$$

Proof (i) For all $a \in A$ and all $b \in B$, $d(A, B) \leq d(a, b)$; hence, for all $a \in A$, $d(A, B) \leq d(a, B)$ and so $d(A, B) \leq \inf \{d(a, B) : a \in A\}$. Now let $\varepsilon > 0$. There exist $a \in A$ and $b \in B$ such that $d(a, B) \leq d(a, b) < d(A, B) + \varepsilon$; thus

$$\inf \{d(x, B) : x \in A\} < d(A, B) + \varepsilon.$$

As this is true for all $\varepsilon > 0$, (i) follows.
(ii) $d(x, A) = 0$ if, and only if, for all $\varepsilon > 0$, $A \cap B(x, \varepsilon) \neq \emptyset$; and, by Lemma 2.1.16, this is true if, and only if, $x \in \overline{A}$.
(iii) Let $x, y \in X$. Then for all $a \in A$,

$$d(x, A) - d(x, y) \leq d(x, a) - d(x, y) \leq d(y, a).$$

Hence $d(x, A) - d(x, y) \leq d(y, A)$. Interchange of x and y shows that $d(y, A) - d(y, x) \leq d(x, A)$, and (iii) follows. □

The notion of uniform convergence for sequences of real-valued functions, introduced in Sect. 1.7, may be developed further to include functions with range in a general metric space.

Definition 2.1.41 Let S be a non-empty set, let (X, d) be a metric space and, for each $n \in \mathbf{N}$, let $f_n : S \to X$. The sequence (f_n) is said to **converge pointwise on** S if there is a function $f : S \to X$ such that for each $s \in S$,

$$\lim_{n \to \infty} d(f_n(s), f(s)) = 0;$$

it is said to **converge uniformly on** S if there is a function $f : S \to X$ such that

$$\lim_{n\to\infty} \sup_S d(f_n(s), f(s)) = 0.$$

Evidently uniform convergence on S implies pointwise convergence on S; apart from special cases, the converse is false. In Sect. 1.7 it was observed that the limit of a uniformly convergent sequence of continuous real-valued functions defined on a subspace of $\mathbf{R}$ is itself a continuous function. This observation carries over to the setting of a general metric space: loosely speaking, uniform convergence preserves continuity. Henceforth, it will be convenient to use the symbol $C(X, Y)$ to denote the family of all continuous functions from a metric space X to a metric space Y; $C(X, \mathbf{R})$ ($\mathbf{R}$ being given the usual metric) may be abbreviated to $C(X)$.

Theorem 2.1.42 *Let X and Y be metric spaces and let $a \in X$. For each $n \in \mathbf{N}$ let $f_n : X \to Y$ be continuous at a, and suppose that the sequence (f_n) converges uniformly on X to $f : X \to Y$. Then f is continuous at a. In particular, if each $f_n \in C(X, Y)$, then $f \in C(X, Y)$.*

Proof This is an obvious modification of that of Theorem 1.7.7. □

Our goal in this subsection is to prove that a continuous real-valued function on a closed subspace of a metric space X has a continuous real-valued extension to all of X. The lemma which follows, usually referred to as Urysohn's lemma , is a special case of this result. Framed in the metric space context adequate for our purposes, it has an elementary proof: Urysohn established it in a more general setting.

Lemma 2.1.43 *Let A and B be disjoint closed subsets of a metric space (X, d). Then there exists a continuous map $f : X \to \mathbf{R}$ such that $f(x) = 1$ $(x \in A)$, $f(x) = -1$ $(x \in B)$ and $|f(x)| \leq 1$ on X.*

Proof If either A or B is empty, then a suitable constant map may be chosen for f. Suppose that neither A nor B is empty and, adopting the notation of Lemma 2.1.40, define $f : X \to \mathbf{R}$ by $f(x) = \{d(x, B) - d(x, A)\}/\{d(x, B) + d(x, A)\}$. Part (ii) of Lemma 2.1.40 shows that, for all $x \in X$, $d(x, B) + d(x, A) > 0$ and so f is well-defined; part (iii) shows that the maps $x \longmapsto d(x, A)$ and $x \longmapsto d(x, B)$ are continuous and therefore (see also Example 2.1.30 (ii)) f is continuous. That f has the remaining properties is clear; indeed, on $X\backslash(A \cup B)$, $|f| < 1$. □

With this lemma at our disposal we give the promised extension theorem, which is due to Tietze.

Theorem 2.1.44 (Tietze's extension theorem) *Let A be a non-empty closed subset of a metric space X and let $f : A \to \mathbf{R}$ be continuous. Then there exists a continuous map $g : X \to \mathbf{R}$ such that $g(x) = f(x)$ for all $x \in A$.*

Proof Let $\phi : \mathbf{R} \to (-1, 1)$ be defined by

$$\phi(x) = (1 + |x|)^{-1}x.$$

Let $h = \phi \circ f$; plainly, $|h| < 1$ on A. We begin by proving that the bounded map h has a continuous extension to X. Let E, F be the disjoint closed sets given by

$$E = \{x \in A : h(x) \leq -1/3\}, F = \{x \in A : h(x) \geq 1/3\}.$$

By the Urysohn Lemma, there exists a continuous map $u_1 : X \to \mathbf{R}$ such that

$$u_1(x) = -1/3 \ (x \in E), u_1(x) = 1/3 \ (x \in F)$$

and $|u_1(x)| \leq 1/3 \ \ (x \in X)$. Note that

$$|h(x) - u_1(x)| < 2/3 \ \ (x \in A).$$

With h replaced by $\frac{3}{2}(h - u_1)$, repetition of the above argument shows that there exists a continuous map $u_2 : X \to \mathbf{R}$ such that

$$|u_2(x)| \leq \frac{2}{3^2} \ (x \in X)$$

and

$$|h(x) - u_1(x) - u_2(x)| < (2/3)^2 \ (x \in A).$$

Inductively, it follows that there is a sequence (u_n) of continuous real-valued functions on X such that

$$|u_n(x)| \leq \frac{2^{n-1}}{3^n} \ (x \in X; n \in \mathbf{N}) \tag{2.1.1}$$

and

$$\left| h(x) - \sum_{k=1}^{n} u_k(x) \right| < (2/3)^n \ (x \in A; n \in \mathbf{N}). \tag{2.1.2}$$

Use of (2.1.1), Theorem 1.7.5 (the Weierstrass $M-$test) and Theorem 2.1.42 shows that $\sum u_n$ converges to a continuous function on X, say u. Then, for all $x \in X$,

$$|u(x)| \leq \sum_{n=1}^{\infty} |u_n(x)| \leq \sum_{n=1}^{\infty} \frac{2^{n-1}}{3^n} = 1;$$

also, from (2.1.2), $u \mid_A = h$.

The map u is a continuous extension of h to X, but this extension may assume the values -1 or 1. A continuous extension eliminating this possibility is constructed next. Let

$$B = \{x \in X : u(x) \in \{-1, 1\}\}.$$

Plainly, B is closed and $A \cap B = \emptyset$. By Urysohn's Lemma, there exists a continuous map $\psi : X \to \mathbf{R}$ such that

$$\psi(x) = 1 \ (x \in A), \ \psi(x) = 0 \ (x \in B)$$

and $0 \leq \psi \leq 1$ on $X \backslash (A \cup B)$. Let $v : X \to \mathbf{R}$ be defined by

$$v(x) = \psi(x)u(x) \ \ (x \in X).$$

Evidently, $v\mid_A = u\mid_A = h$, and v is a continuous extension of h with values in $(-1, 1)$.

Lastly, the conclusion of the theorem is immediate on setting $g = \phi^{-1} \circ v$. Incidentally, it can be shown (see [5], (4.5.1)) that there is an extension g with the property that

$$\sup_{x \in X} g(x) = \sup_{y \in A} f(y), \ \inf_{x \in X} g(x) = \inf_{y \in A} f(y).$$

□

Exercise 2.1.45

1. Let $1 < p < \infty$ and define p' by $\frac{1}{p} + \frac{1}{p'} = 1$. Let $f : [0, \infty) \to \mathbf{R}$ be given by $f(t) = \frac{t^p}{p} + \frac{1}{p'} - t$. Show that the minimum of f on $[0, \infty)$ is attained only at $t = 1$ and that the minimum value is 0. Hence show that for all $a, b \in [0, \infty)$,

$$ab \leq \frac{a^p}{p} + \frac{b^{p'}}{p'},$$

with equality if, and only if, $a = b^{1/(p-1)}$.
Deduce Hölder's inequality: for every $x = (x_1, \ldots x_n), y = (y_1, \ldots, y_n) \in \mathbf{R}^n$,

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} \left(\sum_{k=1}^{n} |y_k|^{p'} \right)^{1/p'}.$$

(The case $p = 2$ is Schwarz's inequality.) Use this to prove Minkowski's inequality:

$$\left(\sum_{k=1}^{n} |x_k + y_k|^p \right)^{1/p} \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{1/p} + \left(\sum_{k=1}^{n} |y_k|^p \right)^{1/p}.$$

Hence show that $(\mathbf{R}^n, d_p)$ is a metric space, where

$$d_p(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^p \right)^{1/p}.$$

2. Let $a, b \in \mathbf{R}$, suppose that $a < b$ and put $I = [a, b]$; let $1 \leq p < \infty$. Prove that $(C(I), d)$ is a metric space, where

$$d(f, g) = \left(\int_a^b |f(t) - g(t)|^p \, dt \right)^{1/p} \qquad (f, g \in C(I)).$$

3. Let $p \in [1, \infty)$ and set

$$\ell_p = \left\{ x = (x_i)_{i \in \mathbf{N}} : x_i \in \mathbf{R} \text{ for all } i \in \mathbf{N},\ \sum_1^\infty |x_i|^p < \infty \right\},$$

$$d_p(x, y) = \left(\sum_{i=1}^\infty |x_i - y_i|^p \right)^{1/p} \quad (x, y \in \ell_p).$$

Prove that (ℓ_p, d_p) is a metric space.

4. Let S be the set of all sequences of real numbers and define d by

$$d(x, y) = \sum_{n=1}^\infty \frac{|x_n - y_n|}{2^n \left[1 + |x_n - y_n|\right]} \quad (x = (x_n), y = (y_n) \in S).$$

Show that (S, d) is a metric space.
[Hint: $t \longmapsto t/(1+t)$ is an increasing function on $[0, \infty)$.]

5. Let p be a prime number. Given any distinct integers m, n, let $t = t(m, n)$ be the unique integer such that

$$m - n = kp^t$$

for some integer k not divisible by p. Define $d : \mathbf{Z} \times \mathbf{Z} \to \mathbf{R}$ by

$$d(m, n) = \begin{cases} 1/p^t & \text{if } m \neq n, \\ 0 & \text{if } m = n. \end{cases}$$

Prove that for all distinct $a, b, c \in \mathbf{Z}$,

$$t(a, c) \geq \min \{t(a, b), t(b, c)\},$$

and hence show that $(\mathbf{Z}, d)$ is a metric space.

6. Determine whether the following subsets of $\mathbf{R}$ (endowed with the usual metric) are open, closed or neither open nor closed:

$$\text{(i) } \mathbf{N}, \quad \text{(ii) } \left\{ \frac{1}{n} : n \in \mathbf{N} \right\}, \quad \text{(iii) } \mathbf{Q}, \quad \text{(iv) } \left\{ (-1)^n \left(1 + \frac{1}{n}\right) : n \in \mathbf{N} \right\}.$$

7. Show that each of the following sets is an open subset of $\mathbf{R}^2$, endowed with the Euclidean metric:

   (i) $\left\{(x, y) : x^2 + y^2 < 1,\ x > 0,\ y > 0\right\}$,
   (ii) $\{(x, y) : x + y \neq 0\}$,
   (iii) $\{(x, y) : xy \neq 1\}$.

   Is $\{(x, 0) : 0 < x < 1\}$ an open subset of $\mathbf{R}^2$?
8. Show that each subset of a discrete metric space is open and closed.
9. Let S be the set of all sequences of real numbers; given any $x = (x_n)$ and $y = (y_n)$ in S, with $x \neq y$, let $k(x, y)$ be the smallest integer n such that $x_n \neq y_n$. Define $d : S \times S \to \mathbf{R}$ by

$$d(x, y) = \begin{cases} 1/k(x, y) & \text{if} \quad x \neq y, \\ 0 & \text{if} \quad x = y. \end{cases}$$

   Prove that for all $x, y, z \in S$,

$$d(x, y) \leq \max\{d(x, z), d(z, y)\},$$

   and hence show that (S, d) is a metric space.
10. Let (X, d) be a metric space and suppose that for all $x, y, z \in X$,

$$d(x, y) \leq \max\{d(x, z), d(y, z)\}.$$

   Prove that if $d(x, z) \neq d(y, z)$, then $d(x, y) = \max\{d(x, z), d(y, z)\}$. Show also that if $x \in X$ and $r > 0$, then $B(x, r) = B(y, r)$ for all $y \in B(x, r)$. Prove that if two open balls in (X, d) intersect, then one is contained in the other. Show that for all $x \in X$ and all $r > 0$, $B(x, r)$ is closed and $\{y \in X : d(x, y) \leq r\}$ is open.
11. Let X be a non-empty set and $d : X \times X \to \mathbf{R}$ a mapping such that

   (i) $d(x, y) = 0$ if, and only if, $x = y$;
   (ii) $d(x, z) \leq d(x, y) + d(z, y)$ for all $x, y, z \in X$.

   Prove that (X, d) is a metric space.
12. Let d_1, d_2 be two metrics on a non-empty set X, and suppose that there are positive constants α, β such that for all $x, y \in X$,

$$\alpha d_1(x, y) \leq d_2(x, y) \leq \beta d_1(x, y).$$

   Prove that the metric spaces (X, d_1) and (X, d_2) have the same open sets. Deduce that each of the metrics d_p of Example 2.1.2 (iii) generates the same family of open subsets of $\mathbf{R}^n$.

13. Show that for all subsets A and B of a metric space X,

$$\overset{o}{A} \cap \overset{o}{B} = \overset{o}{\widehat{A \cap B}}, \quad \overline{A} \cup \overline{B} = \overline{A \cup B}.$$

Show by means of examples that, in general,

$$\overset{o}{A} \cup \overset{o}{B} \neq \overset{o}{\widehat{A \cup B}}, \quad \overline{A} \cap \overline{B} \neq \overline{A \cap B}.$$

Find the closure and interior of the subset D of $\mathbf{R}^3$ (with the Euclidean metric) defined by

$$D = \left\{(x, y, z) \in \mathbf{R}^3 : \cosh(x + yz) \geq 2\right\}.$$

14. Determine the interiors and closures of the following subsets of $\mathbf{R}^2$ (with the Euclidean metric):
(i) $\{(x, y) : 0 < x \leq y < 1\}$, (ii) $\{(x, 0) : 0 < x < 1\}$, (iii) $\{(x, y) : x, y \in \mathbf{Q}\}$.
15. Let S be the subset of $[0, 1]$ (with the usual metric) consisting of all those real numbers which have a decimal representation of the form

$$\sum_{n=1}^{\infty} \frac{a_n}{10^n},$$

where $a_n \in \{0, 1\}$ for all $n \in \mathbf{N}$. By consideration of any $y \in [0, 1]\backslash S$ and the first digit in the decimal representation of y which is not 0 or 1, find the closure of S.
16. By consideration of a discrete metric space, show that a closed ball in a metric space need not be the closure of the open ball with the same centre and the same radius.
17. Let (X, d) be a metric space, let Y be a non-empty subset of X and let d_Y be the restriction of d to $Y \times Y$.

(i) Show that A is a closed subset of (Y, d_Y) if and only if there is a closed subset B of (X, d) such that $A = B \cap Y$.
(ii) Let Y be a closed subset of (X, d) and $S \subset Y$; let $cl_Y(S)$ $(cl_X(S))$ denote the closure of S in (Y, d_Y) $((X, d))$. Show that $cl_Y(S) = cl_X(S)$.

18. Let $f : \mathbf{R}^2 \to \mathbf{R}$ be defined by

$$f(x, y) = \begin{cases} xy/(x^2 + y^2) & \text{if } (x, y) \neq (0, 0), \\ 0 & \text{if } (x, y) = (0, 0); \end{cases}$$

$\mathbf{R}^2$ and $\mathbf{R}$ are each supposed to be equipped with the appropriate Euclidean metric. Show that f is continuous at each point of $\mathbf{R}^2\backslash\{(0, 0)\}$, and that it is not continuous at $(0, 0)$.

19. Let X be a metric space, let $f, g : X \to \mathbf{R}$ be continuous and define $h : X \to \mathbf{R}^2$ by $h(x) = (f(x), g(x))$ $(x \in X)$. Given that $\mathbf{R}^2$ is endowed with the Euclidean metric, show that h is continuous on X.
20. Discuss the continuity of the map $f : \mathbf{R}^2 \to \mathbf{R}^2$ ($\mathbf{R}^2$ is equipped with the Euclidean metric) defined by

$$f(x, y) = \begin{cases} \left(\dfrac{x^2 - y^2}{x^2 + y^2}, \dfrac{(x^2 - y^2)^2}{x^2 + y^2} \right) & \text{if } (x, y) \neq (0, 0), \\ (0, 0) & \text{if } (x, y) = (0, 0). \end{cases}$$

21. Show that $S := \{(x, y) : x^2 - y^2 + 2xy < 0\}$ is an open subset of $\mathbf{R}^2$ (equipped with the Euclidean metric).
22. Let A and B be non-empty subsets of a metric space (X, d). Prove that

    (i) A is bounded if, and only if, there exist $x \in X$ and $r > 0$ such that $A \subset B(x, r)$;
    (ii) $A \subset B$ implies that $\operatorname{diam}(A) \leq \operatorname{diam}(B)$;
    (iii) $\operatorname{diam}(A) = 0$ if, and only if, for some $x \in X$, $A = \{x\}$;
    (iv) if $a \in A$ and $b \in B$, then

    $$\operatorname{diam}(A \cup B) \leq \operatorname{diam}(A) + \operatorname{diam}(B) + d(a, b);$$

    (v) if A and B are bounded, then $A \cup B$ is bounded; further, a finite union of bounded subsets of X is bounded.

23. Let A be a non-empty set of real numbers which is bounded above and let $a = \sup A$. Prove that $a \in \overline{A}$.
24. Let A and B be closed, disjoint subsets of a metric space X. Show that there are open, disjoint subsets U and V (of X) such that $A \subset U$ and $B \subset V$. [Hint: Urysohn's lemma.]

2.2 Complete Metric Spaces

An important property of real numbers is that every Cauchy sequence in $\mathbf{R}$ converges to a point of $\mathbf{R}$. We distinguish a class of metric spaces in which the same kind of property holds. These spaces, the *complete* spaces, are of the utmost theoretical and practical importance.

Definition 2.2.1 Let (X, d) be a metric space. A sequence (x_n) in X is called a **Cauchy sequence** if given any $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that $d(x_m, x_n) < \varepsilon$ whenever $m, n \geq N$; equivalently, $\operatorname{diam}\{x_m : m \geq n\} \to 0$ as $n \to \infty$. Loosely, these conditions may be written $d(x_m, x_n) \to 0$ as $m, n \to \infty$. The space X is said to be **complete** if given any Cauchy sequence (x_n) in X , there exists $x \in X$ such that $x_n \to x$ as $n \to \infty$.

Example 2.2.2

(i) **R**, with the usual metric, is complete: this was our prototype.

(ii) $\mathbf{R}^n$, with the usual (Euclidean) metric d_2, is complete. To prove this, let $(x^{(m)})$ be a Cauchy sequence in $\mathbf{R}^n$, with $x^{(m)} = (x_1^{(m)}, \ldots, x_n^{(m)})$. For each $j \in \{1, \ldots, n\}$,

$$\left|x_j^{(m)} - x_j^{(p)}\right| \leq \left(\sum_{k=1}^{n} \left|x_k^{(m)} - x_k^{(p)}\right|^2\right)^{1/2} = d_2(x^{(m)}, x^{(p)}) \to 0$$

as $m, p \to \infty$; that is, $(x_j^{(m)})_{m \in \mathbf{N}}$ is a Cauchy sequence in **R** and hence converges, to $x_j \in \mathbf{R}$, say. Put $x = (x_1, \ldots, x_n) \in \mathbf{R}^n$. Then $d_2(x^{(m)}, x) = \left(\sum_{k=1}^{n} \left|x_k^{(m)} - x_k\right|^2\right)^{1/2} \to 0$ as $m \to \infty$: $\mathbf{R}^n$ is complete.

(iii) **Q**, the set of all rationals, with the usual metric inherited from **R**, is not complete: $\left(\left(1 + \frac{1}{n}\right)^n\right)_{n \in \mathbf{N}}$ is a Cauchy sequence in **Q** which does not converge to an element of **Q**.

(iv) The open interval $(0, 2)$, with the usual metric inherited from **R**, is not complete: $\left(\frac{1}{n}\right)_{n \in \mathbf{N}}$ is a Cauchy sequence in $(0, 2)$ which fails to converge to an element of $(0, 2)$.

(v) Let $I = [0, 1]$, take $X = C(I)$ (the set of all continuous, real-valued functions on I) and define a metric d on $C(I)$ by

$$d(f, g) = \int_0^1 |f(t) - g(t)|\, dt.$$

Then $(C(I), d)$ is not complete. To establish this, consider the sequence $(f_n)_{n \geq 2}$, where

$$f_n(t) = \begin{cases} 0, & \text{if } 0 \leq t \leq \frac{1}{2}, \\ n(t - \frac{1}{2}), & \text{if } \frac{1}{2} < t \leq \frac{1}{2} + \frac{1}{n}, \\ 1, & \text{if } \frac{1}{2} + \frac{1}{n} \leq t \leq 1. \end{cases}$$

Since $d(f_n, f_m) = \frac{1}{2}\left|m^{-1} - n^{-1}\right| \to 0$ as $m, n \to \infty$, (f_n) is a Cauchy sequence. Suppose that there is a function $f \in C(I)$ such that $d(f_n, f) \to 0$ as $n \to \infty$. Then $\int_0^{1/2} |f_n(t) - f(t)|\, dt \leq d(f_n, f) \to 0$ as $n \to \infty$; thus $\int_0^{1/2} |f(t)|\, dt = 0$, and hence $f(t) = 0$ for all $t \in \left[0, \frac{1}{2}\right]$. Now let $\varepsilon \in \left(0, \frac{1}{2}\right)$. Since

$$\int_{\frac{1}{2}+\varepsilon}^{1} |f_n(t) - f(t)|\, dt \leq d(f, f_n) \to 0 \text{ as } n \to \infty,$$

and for all large enough $n, f_n(t) = 1$ on $\left[\frac{1}{2} + \varepsilon, 1\right]$, we see that

$$\int_{\frac{1}{2}+\varepsilon}^{1} |f(t) - 1|\, dt = 0,$$

so that $f(t) = 1$ for all $t \in \left[\frac{1}{2} + \varepsilon, 1\right]$. As this holds for all $\varepsilon \in (0, \frac{1}{2})$, it follows that $f(t) = 1$ for all $t \in \left(\frac{1}{2}, 1\right]$, which implies that f is discontinuous at $t = \frac{1}{2}$. This contradiction shows that $(C(I), d)$ is not complete.

(vi) Let $p \in [1, \infty)$, let

$$\ell_p = \left\{ x = (x_i)_{i \in \mathbf{N}} : x_i \in \mathbf{R} \text{ for all } i \in \mathbf{N},\ \sum_1^\infty |x_i|^p < \infty \right\},$$

and let

$$d_p(x, y) = \left(\sum_1^\infty |x_i - y_i|^p \right)^{1/p},$$

where $x = (x_i)$, $y = (y_i) \in \ell_p$. Then (ℓ_p, d_p) is complete. To prove this, let $(x^n)_{n \in \mathbf{N}}$ be a Cauchy sequence in ℓ_p, where $x^n = \left(x_i^n\right)_{i \in \mathbf{N}}$. For each $i \in \mathbf{N}$,

$$\left|x_i^m - x_i^n\right| \le d_p(x^m, x^n)$$

and hence $\left(x_i^n\right)_{n \in \mathbf{N}}$ is a Cauchy sequence in $\mathbf{R}$. Using the completeness of $\mathbf{R}$, let $x_i = \lim_{n \to \infty} x_i^n$ and put $x = (x_i)_{i \in \mathbf{N}}$. It remains to show that $x \in \ell_p$ and that $x^n \to x$ in ℓ_p.

Let $\varepsilon > 0$. There exists $N \in \mathbf{N}$ such that $d_p(x^m, x^n) < \varepsilon$ if $m, n \ge N$. Thus for each $k \in \mathbf{N}$,

$$\sum_{i=1}^k \left|x_i^m - x_i^n\right|^p < \varepsilon^p \text{ if } m, n \ge N;$$

thus (letting $m \to \infty$),

$$\sum_{i=1}^k \left|x_i - x_i^n\right|^p \le \varepsilon^p \text{ if } n \ge N. \tag{2.2.1}$$

Use of (2.2.1) in conjunction with Minkowski's inequality shows that, for each $k \in \mathbf{N}$,

$$\left(\sum_{i=1}^{k}|x_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{k}\left|x_i - x_i^N\right|^p\right)^{1/p} + \left(\sum_{i=1}^{k}\left|x_i^N\right|^p\right)^{1/p}$$

$$\leq \varepsilon + \left(\sum_{i=1}^{\infty}\left|x_i^N\right|^p\right)^{1/p}.$$

Hence $x \in \ell_p$. Further use of (2.2.1) shows that

$$d_p(x^n, x) = \left(\sum_{i=1}^{\infty}\left|x_i - x_i^n\right|^p\right)^{1/p} \leq \varepsilon.$$

Hence $x^n \to x$.

To make the interval $(0, 2)$ of Example 2.2.2 (iv) above into a complete space all we have to do is to adjoin the two points 0 and 2; the space $\mathbf{Q}$ of Example 2.2.2 (iii) may be 'completed' by adjoining all irrationals. These two examples illustrate the general principle, examined later, that any incomplete space may be enlarged so as to make it into a complete space.

The following result gives a useful characterisation of complete spaces; it uses the so-called *Cantor intersection property*: a metric space (X, d) is said to have this property if whenever (A_n) is a sequence of non-empty, closed, bounded subsets of X such that $A_{n+1} \subset A_n$ for all $n \in \mathbf{N}$ and $\lim_{n\to\infty} \operatorname{diam} A_n = 0$, then $\bigcap_{n=1}^{\infty} A_n$ has exactly one point.

Theorem 2.2.3 (Cantor's characterisation of completeness) *A metric space (X, d) is complete if, and only if, X has the Cantor intersection property.*

Proof First suppose that X is complete, and let (A_n) be a sequence of non-empty, closed bounded subsets of X such that $A_{n+1} \subset A_n$ for all $n \in \mathbf{N}$ and $\lim_{n\to\infty} \operatorname{diam} A_n = 0$; let (x_n) be a sequence such that for all $n \in \mathbf{N}$, $x_n \in A_n$. If $m \geq n$, then $x_m \in A_n$ and $\operatorname{diam}\{x_m : m \geq n\} \leq \operatorname{diam} A_n \to 0$ as $n \to \infty$. Hence (x_n) is a Cauchy sequence and $x := \lim_{n\to\infty} x_n \in \overline{A_k} = A_k$ for all $k \in \mathbf{N}$; so $x \in \bigcap\limits_{1}^{\infty} A_k$. If $y \in \bigcap_1^{\infty} A_k$, then $d(x, y) \leq \operatorname{diam} A_n \to 0$ as $n \to \infty$; hence $y = x$. It follows that $\bigcap_1^{\infty} A_n = \{x\}$.

Conversely, suppose that X has the Cantor intersection property and let (x_n) be a Cauchy sequence in X. Let A_n be the closure of $\{x_m : m \geq n\}$ $(n \in \mathbf{N})$; then $A_{n+1} \subset A_n$ and $\operatorname{diam} A_n \to 0$ as $n \to \infty$. Thus there exists a unique $x \in X$ such that $x \in \bigcap\limits_{1}^{\infty} A_n$; and $d(x, x_m) \leq \operatorname{diam} A_m \to 0$ as $m \to \infty$, that is, $x_m \to x$. Hence X is complete. $\square$

Augmenting the complete metric spaces already described, we now introduce further examples each of which provides a suitable context for specific problems.

Definition 2.2.4 Let S be a non-empty set and let $\mathscr{B}(S)$ be the family of all bounded, real-valued functions on S. The **uniform metric** d_∞ on $\mathscr{B}(S)$ is given by

$$d_\infty(f, g) = \sup\{|f(s) - g(s)| : s \in S\} \ (f, g \in \mathscr{B}(S)).$$

If S is a metric space, $\mathscr{C}(S)$ stands for the family of all continuous, bounded, real-valued functions on S; the restriction of d_∞ to $\mathscr{C}(S)$ is again denoted by d_∞.

Note that $d_\infty(f_n, f) \to 0$ if, and only if, (f_n) converges to f uniformly on S.

The arguments needed for the proofs of the next two theorems are essentially those given in Sect. 1.7, but we give the details for the convenience of the reader.

Theorem 2.2.5 *The metric space* $(\mathscr{B}(S), d_\infty)$ *is complete.*

Proof Let (f_n) be a Cauchy sequence in $\mathscr{B}(S)$. Then given any $\varepsilon > 0$, there exists $N \in \mathbf{N}$ such that $d_\infty(f_n, f_m) < \varepsilon$ if $m, n \geq N$; and so for each $s \in S$, $|f_n(s) - f_m(s)| < \varepsilon$ if $m, n \geq N$. Thus for each $s \in S$, $(f_n(s))$ is a Cauchy sequence in $\mathbf{R}$ and hence converges, to $f(s)$, say. We thus have a map $f : S \to \mathbf{R}$, where $f(s) = \lim_{n\to\infty} f_n(s)$ for all $s \in S$. To complete the proof we must show that $f \in \mathscr{B}(S)$ and $d(f_n, f) \to 0$ as $n \to \infty$. As above, we see that for all $s \in S$, $|f_n(s) - f_m(s)| < \varepsilon$ if $m, n \geq N$. Let $m \to \infty$: then for all $s \in S$, $|f_n(s) - f(s)| \leq \varepsilon$ if $n \geq N$. Since $|f(s)| \leq |f_N(s)| + \varepsilon$, it follows that $f \in \mathscr{B}(S)$; also we have $d_\infty(f, f_n) \leq \varepsilon$ if $n \geq N$. Hence $f_n \to f$ in $\mathscr{B}(S)$. □

Theorem 2.2.6 *Let* (S, d) *be a metric space. Then* $(\mathscr{C}(S), d_\infty)$ *is complete.*

Proof Let (f_n) be a Cauchy sequence in $\mathscr{C}(S)$. Then (f_n) is a Cauchy sequence in $\mathscr{B}(S)$, and so by Theorem 2.2.5, there exists $f \in \mathscr{B}(S)$ such that $d_\infty(f, f_n) \to 0$ as $n \to \infty$. Let $\varepsilon > 0$. Then there exists $N \in \mathbf{N}$ such that for all $n \geq N$ and all $s \in S$, $|f_n(s) - f(s)| < \varepsilon/3$. Let $s_0 \in S$. Since f_N is continuous at s_0, there exists $\delta > 0$ such that $|f_N(s) - f_N(s_0)| < \varepsilon/3$ if $d(s, s_0) < \delta$. Thus if $d(s, s_0) < \delta$, then

$$|f(s) - f(s_0)| \leq |f(s) - f_N(s)| + |f_N(s) - f_N(s_0)| + |f_N(s_0) - f(s_0)| < \varepsilon.$$

Hence $f \in \mathscr{C}(S)$, and the theorem follows. □

Corollary 2.2.7 *Let* $I = [a, b] \subset \mathbf{R}$. *Then* $C(I) = \mathscr{C}(I)$ *and* $(C(I), d_\infty)$ *is complete.*

Proof That $C(I) = \mathscr{C}(I)$ follows immediately from the fact that every continuous real-valued function on the closed, bounded interval I is bounded. The rest is now clear from Theorem 2.2.6. □

We take up in the next section the question of under what conditions on a metric space S can it be shown that $C(S) = \mathscr{C}(S)$.

Theorem 2.2.8 *Let* $a, b \in \mathbf{R}$ *and* $a < b$. *Then* $(\mathscr{R}[a, b], d_\infty)$ *is complete.*

Proof Let (f_n) be a Cauchy sequence in $\mathscr{R}[a, b]$; it is also a Cauchy sequence in $\mathscr{B}[a, b]$ and so there is an $f \in \mathscr{B}[a, b]$ such that $d_\infty(f, f_n) \to 0$ *as* $n \to \infty$. Evidently (f_n) converges uniformly to f on $[a, b]$. By Theorem 1.7.12 it follows that $f \in \mathscr{R}[a, b]$. □

Theorem 2.2.9 *Let $a, b \in \mathbf{R}$, $a < b$, let $I = [a, b]$ and let $C^1(I)$ denote the family of all continuously differentiable real-valued functions on I. Let $\nu : C(I) \times C(I) \to \mathbf{R}$ be defined by*

$$\nu(f, g) = \sup\{|f(x) - g(x)| : x \in I\} + \sup\left\{\left|f'(x) - g'(x)\right| : x \in I\right\}$$

(*that is,* $\nu(f, g) = d_\infty(f, g) + d_\infty(f', g')$). *Then* $(C^1(I), \nu)$ *is a complete metric space.*

Proof Routine arguments show that ν is a metric on $C^1(I)$. To prove completeness, let (f_n) be a Cauchy sequence in $C^1(I)$. Then (f_n) and (f_n') are Cauchy sequences in the complete space $(C(I), d_\infty)$, and so there exist $f, g \in C(I)$ such that $d_\infty(f_n, f) \to 0$ and $d_\infty(f_n', g) \to 0$ as $n \to \infty$. The result is immediate if we can prove that $f' = g$. However, by Theorem 1.4.4,

$$f_n(x) - f_n(a) = \int_a^x f_n' \ (x \in I, n \in \mathbf{N});$$

and since $\left|\int_a^x (f_n' - g)\right| \le (x - a) d_\infty(f_n', g) \to 0$ as $n \to \infty$, we have that

$$f(x) - f(a) = \lim_{n\to\infty} (f_n(x) - f_n(a)) = \lim_{n\to\infty} \int_a^x f_n' = \int_a^x g \ (x \in I).$$

Thus by Theorem 1.4.9, f is differentiable and $f' = g$. □

Corollary 2.2.10 *Let (f_n) be a sequence in $C^1(I)$ $(I = [a, b])$ such that (f_n') converges uniformly on I and for some $x_0 \in I$, $(f_n(x_0))$ is convergent. Then there exists $f \in C^1(I)$ such that (f_n) converges uniformly on I to f and*

$$f'(x) = \lim_{n\to\infty} f_n'(x) \ (x \in I).$$

Proof It is enough to show that (f_n) is a Cauchy sequence in $(C(I), d_\infty))$, for then (f_n) will be a Cauchy sequence in $(C^1(I), \nu)$ and the result will follow immediately from Theorem 2.2.9. To do this, let $\varepsilon > 0$ and let $N \in \mathbf{N}$ be such that

$$|f_m(x_0) - f_n(x_0)| < \varepsilon/2 \text{ and } d_\infty(f_m', f_n') < \varepsilon/2(b - a) \text{ if } m, n > N.$$

Since for all $x \in I$ we have, by Theorem 1.4.4,

$$|f_m(x) - f_n(x)| \le |f_m(x_0) - f_n(x_0)| + \left| \int_{x_0}^{x} (f'_m - f'_n) \right|$$

$$\le |f_m(x_0) - f_n(x_0)| + (b-a)d_\infty(f'_m, f'_n) < \varepsilon$$

if $m, n > N$. The result follows. $\square$

Returning to the observations concerning the completion of an incomplete metric space following Example 2.2.2, we see that Theorem 2.2.6 leads to the following result.

Theorem 2.2.11 *Let (S, d) be a metric space. Then there is a complete metric space $(\widehat{S}, \widehat{d})$ such that S is isometric to a dense subset S_0 of $\widehat{S}$; that is, a subset S_0 such that $\overline{S}_0 = \widehat{S}$.*

Proof Fix $a \in S$ and for every $p \in S$, define $f_p : S \to \mathbf{R}$ by

$$f_p(x) = d(x, p) - d(x, a).$$

Use of the triangle inequality shows that, for all $x, y \in S$,

$$\left|f_p(x) - f_p(y)\right| \le 2d(x, y), \quad \left|f_p(x)\right| \le d(a, p).$$

Hence $f_p \in \mathscr{C}(S)$. Let $S_0 = \{f_p : p \in S\}$. Since, for all $p, q \in S$,

$$d_\infty(f_p, f_q) = \sup_{x \in S} |d(x, p) - d(x, q)| = d(p, q),$$

the map $p \longmapsto f_p : S \to S_0 \subset \mathscr{C}(S)$ is an isometry of S onto S_0. Let $\widehat{S}$ be the closure of S_0 in $(\mathscr{C}(S), d_\infty)$ and $\widehat{d}$ be the restriction of d_∞ to $\widehat{S} \times \widehat{S}$. Since $\widehat{S}$ is closed in $(\mathscr{C}(S), d_\infty)$, use of Lemma 2.1.24 shows that $\left(\widehat{S}, \widehat{d}\right)$ is complete. Let $\overline{S}_0$ denote the closure of S_0 in $\left(\widehat{S}, \widehat{d}\right)$. As $\overline{S}_0 = \widehat{S}$, S_0 is dense in $\left(\widehat{S}, \widehat{d}\right)$ and S is isometric to it. $\square$

Even though $\widehat{X}$ is not unique in having the property of being complete and having X isometric to a dense subset of $\widehat{X}$—it is only unique up to an isometry—we shall refer to it as *the completion* of X.

Two of the most celebrated results associated with complete metric spaces are the Contraction Mapping Theorem and the Baire Category Theorem. The rest of this section is devoted to establishing these and illustrating their application.

2.2.1 The Contraction Mapping Theorem

This is one of the most useful, yet simple, theorems in mathematics.

Definition 2.2.12 Let (X, d) be a metric space. A map $f : X \to X$ is called a **contraction** if there is a number $\lambda \in [0, 1)$ such that for all $x, y \in X$,

$$d(f(x), f(y)) \leq \lambda d(x, y).$$

Theorem 2.2.13 (Banach's contraction mapping theorem) *Let (X, d) be a complete metric space and let $f : X \to X$ be a contraction. Then there is exactly one point $x \in X$ such that $f(x) = x$; that is, f has exactly one fixed point.*

Proof As f is a contraction, there exists $\lambda \in [0, 1)$ such that $d(f(x), f(y)) \leq \lambda d(x, y)$ for all $x, y \in X$. Let $x_0 \in X$ and define a sequence (x_n) by $x_n = f(x_{n-1})$ $(n \in \mathbf{N})$. Then for each $n \in \mathbf{N}$,

$$d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq \lambda d(x_n, x_{n-1}) \leq \lambda^n d(x_1, x_0).$$

If $m > n$,

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \ldots + d(x_{n+1}, x_n) \\ &\leq (\lambda^{m-1} + \lambda^{m-2} + \ldots + \lambda^n) d(x_1, x_0) \\ &= \frac{(\lambda^n - \lambda^m)}{1 - \lambda} d(x_1, x_0). \end{aligned}$$

It follows that (x_n) is a Cauchy sequence in X and, as X is complete, there exists $x \in X$ such that $x_n \to x$. Thus

$$\begin{aligned} d(x, f(x)) &\leq d(x, x_{n+1}) + d(x_{n+1}, f(x)) \leq d(x, x_{n+1}) + \lambda d(x_n, x) \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$

Hence $f(x) = x$; that is, x is a fixed point of f.

If there exists $y \in X$ such that $f(y) = y$, then

$$d(x, y) = d(f(x), f(y)) \leq \lambda d(x, y);$$

and as $\lambda < 1$, $d(x, y) = 0$. Hence $x = y$, and the proof is complete. □

Note the constructive nature of this proof: no matter what point x_0 of X is chosen, the fixed point x of f is given by the formula

$$x = \lim_{n \to \infty} f^n(x_0).$$

In practical circumstances, approximations to the fixed point may be derived by choosing a convenient point x_0 and determining $f^n(x_0)$ for various values of n.

Corollary 2.2.14 *Let X be a complete metric space and let $f : X \to X$ be such that, for some $k \in \mathbf{N}$, f^k is a contraction. Then f has a unique fixed point.*

Proof By Theorem 2.2.13, there is a unique $x \in X$ such that $f^k(x) = x$. But $f^k(f(x)) = f(f^k(x)) = f(x)$, and so $f(x)$ is a fixed point of f^k. Hence $f(x) = x$. That f has a unique fixed point now follows since evidently any fixed point of f must be a fixed point of f^k, and so must coincide with x. □

We can now give an application of the contraction mapping theorem to the theory of ordinary differential equations.

Theorem 2.2.15 *Let $a, b \in \mathbf{R}$, $a < b$, put $I = [a, b]$, let $f : I \times \mathbf{R} \to \mathbf{R}$ be continuous and suppose there exists $M > 0$ such that for all $x \in I$ and all $y_1, y_2 \in \mathbf{R}$,*

$$|f(x, y_1) - f(x, y_2)| \leq M\, |y_1 - y_2| .$$

Let $c \in \mathbf{R}$. Then there is a unique function $u \in C^1(I)$ such that

$$u'(x) = f(x, u(x))\ (x \in I),\ u(a) = c. \tag{2.2.2}$$

Proof First observe that if $u \in C(I)$, then $t \longmapsto f(t, u(t)) : I \to \mathbf{R}$ is continuous, for if $t \in I$ and (t_n) is a sequence in I with $t_n \to t$, then $(t_n, u(t_n)) \to (t, u(t))$ in $I \times \mathbf{R}$ and so $f(t_n, u(t_n)) \to f(t, u(t))$. Further, by the Fundamental Theorem of Integral Calculus, there is a unique $u \in C^1(I)$ such that (2.2.2) holds if, and only if, the integral equation

$$u(x) = c + \int_a^x f(t, u(t))dt\ (x \in I) \tag{2.2.3}$$

has a unique solution $u \in C(I)$. Define a map $T : C(I) \to C(I)$ by

$$(Tu)(x) = c + \int_a^x f(t, u(t))dt\ (x \in I;\ u \in C(I)).$$

For each $n \in \mathbf{N}_0$, let $P(n)$ be the proposition

$$\left|(T^n u)(x) - (T^n v)(x)\right| \leq (M\,|x - a|)^n d_\infty(u, v)/n!$$

for all $u, v \in C(I)$ and all $x \in I$. (Here T^0 is the identity map of $C(I)$ to itself.)

Plainly $P(0)$ is true; moreover, if $P(n)$ holds for some $n \in \mathbf{N}_0$, then

$$\left|(T^{n+1}u)(x) - (T^{n+1}v)(x)\right| = \left|\int\limits_a^x \left\{f(t, (T^n u)(t)) - f(t, (T^n v)(t))\right\} dt\right|$$

$$\leq M \left|\int\limits_a^x \{(M\,|t-a|)^n d_\infty(u,v)/n!\} dt\right|$$

$$\leq (M\,|x-a|)^{n+1} d_\infty(u,v)/(n+1)!$$

for all $u, v \in C(I)$ and all $x \in I$, and so $P(n+1)$ is true. Hence $P(n)$ is true for all $n \in \mathbf{N}$. It follows that

$$d_\infty(T^n u, T^n v) \leq (M(b-a))^n d_\infty(u,v)/n!$$

for all $u, v \in C(I)$ and all $n \in \mathbf{N}$. Choose $k \in \mathbf{N}$ so large that $(M(b-a))^k/k! < 1$: then T^k is a contraction on the complete space $(C(I), d_\infty)$. By Corollary 2.2.14, there is a unique $u \in C(I)$ such that $Tu = u$; that is, such that

$$u(x) = c + \int_a^x f(t, u(t))dt \ (x \in I).$$

The result follows. □

2.2.2 The Baire Category Theorem

Several formulations of this theorem exist. One of the most accessible is as follows.

Theorem 2.2.16 *Let X be a complete metric space and let $(\mathscr{O}_n)$ be a sequence of dense open subsets of X. Then $\cap_{n=1}^\infty \mathscr{O}_n$ is dense in X; that is, $\overline{\cap_{n=1}^\infty \mathscr{O}_n} = X$.*

Proof Suppose that the conclusion is false. Let $U := X \backslash \overline{\cap_{n=1}^\infty \mathscr{O}_n}$: U is open and non-empty. Since $\overline{\mathscr{O}_1} = X$, $U \cap \mathscr{O}_1 \neq \emptyset$. Hence there exists a non-empty open set U_1 such that

$$U_1 \subset \overline{U_1} \subset U \cap \mathscr{O}_1 \text{ and } \operatorname{diam} \overline{U_1} < 1.$$

(U_1 may be taken to be an open ball of suitable radius). Since $\overline{\mathscr{O}_2} = X$, $U_1 \cap \mathscr{O}_2 \neq \emptyset$ and so there is a non-empty open set U_2 such that

$$U_2 \subset \overline{U_2} \subset U_1 \cap \mathscr{O}_2 \text{ and } \operatorname{diam} \overline{U_2} < 2^{-1}.$$

Continuing in this way we see that there exists a sequence (U_n) of non-empty open subsets of X such that, for all $n \in \mathbf{N}$,

$$U_n \subset \overline{U_n} \subset U_{n-1} \cap \mathscr{O}_n \text{ and } \operatorname{diam} \overline{U_n} < n^{-1}.$$

(Here $U_0 := U$.) Thus Cantor's characterisation of completeness (Theorem 2.2.3) shows that for some $x \in X$,

$$\{x\} = \cap_{n=1}^{\infty} \overline{U_n}.$$

Since $\overline{U_1} \subset U$ and, for all $n \in \mathbf{N}$, $\overline{U_n} \subset \mathscr{O}_n$, it follows that $x \in U \cap (\cap_{n=1}^{\infty} \mathscr{O}_n)$, a contradiction. □

Taking complements, and recalling that ${}^c(\overset{o}{A}) = \overline{{}^cA}$ whenever $A \subset X$, we immediately obtain the equivalent result:

Theorem 2.2.17 *Let X be a complete metric space and let (F_n) be a sequence of closed subsets of X, each with empty interior. Then $\cup_{n=1}^{\infty} F_n$ has empty interior.*

Breaking the theoretical development for a moment, we use this last result to give a striking demonstration of the existence of a continuous nowhere-differentiable function.

Theorem 2.2.18 *Let I be the closed interval* $[0, 1]$. *Then there exists an element of $C(I)$ which is not differentiable at any point of I.*

Proof For each $n \in \mathbf{N}$ put

$$M_n = \left\{ f \in C([0,2]) : \text{ for some } x_0 \in I, \sup_{0<h<1} \frac{|f(x_0+h) - f(x_0)|}{h} \leq n \right\}.$$

We claim that each M_n is closed in $C([0, 2])$. To prove this, let $f \in \overline{M_n}$ and let (f_k) be a sequence in M_n that converges to f. For each $k \in \mathbf{N}$, there exists $x_k \in I$ with

$$|f_k(x_k + h) - f_k(x_k)| \leq nh \text{ if } 0 < h < 1.$$

As the bounded sequence (x_k) contains a convergent subsequence, we may and shall assume, without loss of generality, that $x_k \to x_0 \in I$. For all $k \in \mathbf{N}$,

$$\begin{aligned} |f(x_0+h) - f(x_0)| \leq{} & |f(x_0+h) - f(x_k+h)| + |f(x_k+h) - f_k(x_k+h)| \\ & + |f_k(x_k+h) - f_k(x_k)| + |f_k(x_k) - f(x_k)| \\ & + |f(x_k) - f(x_0)|, \end{aligned}$$

and using the fact that $|f_k(x_k + h) - f_k(x_k)| \leq nh$ we see, on letting $k \to \infty$, that

$$|f(x_0+h)-f(x_0)| \leq nh \text{ if } 0 < h < 1.$$

It follows that $f \in M_n$ and our claim is justified.

Next we claim that each M_n has empty interior. For let g be a piecewise-linear continuous function on $[0, 2]$, so that the graph of g consists of a finite number of straight-line segments; let M be the maximum absolute value of the gradients of these segments. Given $\varepsilon > 0$, choose $m \in \mathbf{N}$ so that $m\varepsilon > n + M$, define $\phi : \mathbf{R} \to I$ by

$$\phi(x) = \min\{x - [x], [x] + 1 - x\}, \ x \in \mathbf{R}$$

(here $[x]$ denotes the integer part of x; $\phi(x)$ is simply the distance of x from the nearest integer), and put

$$F(x) = g(x) + \varepsilon\phi(mx), \ x \in I.$$

Then if $x \in I$ and $0 < h < 1$,

$$\begin{aligned}|F(x+h) - F(x)| &= |g(x+h) - g(x) + \varepsilon\{\phi(m(x+h)) - \phi(mx)\}| \\ &\geq \varepsilon mh - |g(x+h) - g(x)| \\ &\geq \varepsilon mh - Mh > nh.\end{aligned}$$

Hence $F \in C([0, 2])\backslash M_n$. Moreover, $d_\infty(g, F) = \varepsilon$. Assuming for the moment that the set of all piecewise-linear continuous functions is dense in $C([0, 2])$, our analysis shows that any $f \in C([0, 2])$ may be approximated arbitrarily closely in $C([0, 2])$ by an element of $C([0, 2])\backslash M_n$, so that the interior of M_n must be empty, as claimed. Since $C([0, 2])$ is complete, it follows from Theorem 2.2.17 that $C([0, 2])\backslash \cup_{n=1}^{\infty} M_n$ is non-empty; and as every function in $C([0, 2])$ that is differentiable at some point of I must lie in some M_n, the theorem follows.

All that remains is to establish the density of the piecewise-linear continuous functions in $C([0, 2])$. Let $f \in C([0, 2])$ and let $\varepsilon > 0$. Since f is uniformly continuous on $[0, 2]$, there is a partition $P = \{0, 2/m, 4/m, \ldots, 2\}$ of $[0, 2]$ such that for $j = 1, \ldots, m$ we have $\operatorname{osc}(f, [(2j-1)/m, 2j/m]) < \varepsilon$ (see Exercise 1.1.10 /2). Define ψ on $[0, 2]$ by

$$\psi(x) = \frac{m}{2}\left\{\left(x - \frac{2(j-1)}{m}\right)f(2j/m) + \left(\frac{2j}{m} - x\right)f(2(j-1)/m)\right\}$$

if $2(j-1)/m \leq x \leq 2j/m, \ j = 1, \ldots, m,$
so that ψ is piecewise linear and coincides with f at the points of the partition. Evidently $d_\infty(f, \psi) < \varepsilon$, and the density follows. $\square$

Alternative formulations of the Baire theorem demand a little preparation.

Definition 2.2.19 Let A and B be subsets of a metric space X. Then A is said to be **dense in** B if $B \subset \overline{A}$; it is **everywhere dense** (or simply **dense**) if it is dense in X; and it is **nowhere dense** (or **rare**) if it is not dense in any non-empty open subset of X, or equivalently, if its closure contains no interior points.

Remark 2.2.20

(i) Plainly, a subset of a nowhere dense set is nowhere dense; also, the closure of a nowhere dense set is itself nowhere dense.
(ii) A closed set is nowhere dense if, and only if, it coincides with its own boundary. That is, if $A \subset X$ and $A = \overline{A}$, then $\overset{o}{A} = \emptyset$ if, and only if, $\partial A = A \backslash \overset{o}{A} = A$.
(iii) To say that a set is everywhere dense is not the antithesis of saying that it is nowhere dense.

Example 2.2.21

1. Let $x \in X$. Then $\{x\}$ is nowhere dense if, and only if, x is not an isolated point of X : an isolated point y is one having a neighbourhood containing no point of X except y. Note that $\{x\}$ is closed. Thus $\overset{o}{\widehat{\{x\}}} = \emptyset$ if, and only if, every neighbourhood U of x is such that $U \cap^c \{x\} \neq \emptyset$; and this is so if, and only if, x is not an isolated point of X.
2. The boundary of an open (or closed) set in X is always nowhere dense. For let U be an open set in X and let V be an open set in X such that $V \subset \partial U = \overline{U} \cap \overline{{}^cU} = \overline{U} \cap {}^cU$. Then ${}^cV$ is closed and contains U, so that ${}^cV \supset \overline{U} \supset V$, which is possible only if $V = \emptyset$. Note that the boundary of an arbitrary set A in X need not be nowhere dense: for example, A and ${}^cA$ might both be dense, in which case $\partial A = X$.

Lemma 2.2.22 *Let A be a subset of a metric space X. The following three statements are equivalent:*

(i) *A is nowhere dense in X.*
(ii) *${}^cA$ contains an everywhere dense open subset of X.*
(iii) *Each non-empty open set U in X contains a non-empty open set V such that $V \cap A = \emptyset$.*

Proof (i)⇒(ii) Suppose $\overset{o}{\overline{A}} = \emptyset$. Then $\overline{{}^c(\overline{A})} = {}^c(\overset{o}{\overline{A}}) = X$ and so the open set ${}^c(\overline{A})$, which is contained in ${}^cA$, is everywhere dense.
(ii)⇒(iii) Let $\mathscr{O}$ be a dense open subset of X contained in ${}^cA$ and let U be a non-empty open subset of X. Note that $U \cap \mathscr{O} \neq \emptyset$: otherwise, $\mathscr{O}$ is contained in the closed set ${}^cU$, so that $X = \overline{\mathscr{O}} \subset {}^cU$ which implies that $U = \emptyset$. Let $V = U \cap \mathscr{O}$. Then V is non-empty and open, $V \subset U$ and $V \cap A = \emptyset$, since $V \subset \mathscr{O} \subset {}^cA$.

(iii)⇒ (i) To obtain a contradiction, suppose $U := \overset{o}{\overline{A}} \neq \emptyset$. Then there is a non-empty open set $V \subset U$ such that $A \cap V = \emptyset$. Since ${}^cV$ is closed and $A \subset {}^cV$, it follows that $U \subset \overline{A} \subset {}^cV$ and so $\emptyset = U \cap V = V \neq \emptyset$, a contradiction. □

Lemma 2.2.23 *Let X be a metric space.*

(i) *If U and V are each dense open subsets of X, then $U \cap V$ is a dense open subset of X.*
(ii) *If A and B are each nowhere dense subsets of X, then $A \cup B$ is nowhere dense.*

Proof (i) To obtain a contradiction, suppose $\overline{U \cap V} \neq X$. Then $G :=^c \left(\overline{U \cap V}\right)$ is open and non-empty. Since $\overline{U} = X$, $G \cap U$ is open and non-empty: otherwise, ${}^cG \supset U$ and, since ${}^cG$ is closed, ${}^cG \supset \overline{U} = X$, implying that $G = \emptyset$. Since $\overline{V} = X$, similar reasoning shows that $G \cap U \cap V$ is open and non-empty. But this contradicts the fact that $U \cap V \subset {}^cG$. Hence $U \cap V$ is dense in X.
(ii) Since A and B are nowhere dense, each of the sets ${}^c(\overline{A})$, ${}^c(\overline{B})$ is open and dense in X. Thus, using (i), it follows that ${}^c(\overline{A}) \ \cap {}^c(\overline{B}) = {}^c(\overline{A} \cup \overline{B}) = {}^c(\overline{A \cup B})$ is dense in X. Since

$$ {}^c\left(\overset{0}{\overline{A \cup B}}\right) = \overline{{}^c(\overline{A \cup B})} = X, $$

the set $A \cup B$ is nowhere dense. □

Note that (i) and (ii) can obviously be extended to arbitrary **finite** intersections and **finite** unions.

Taken together, Theorems 2.2.16 and 2.2.17 extend the last Lemma to countably infinite families of sets. But the extension comes at a price. Recall that a countable intersection of open sets need not be open, and a countable union of closed sets need not be closed. The theorems demand a stronger hypothesis, namely the completeness of X, and support a weaker conclusion than that of the Lemma. To illustrate by example, let $X = \mathbb{R}$ and $A_n = \{x_n\}$, where the sequence $(x_n)_{n \in \mathbb{N}}$ is an enumeration of the rationals. Then $\mathbb{Q} = \cup_{n=1}^{\infty} A_n$ is a countably infinite union of nowhere dense sets and its interior is empty. However, it is not nowhere dense; indeed, it is everywhere dense.

Definition 2.2.24 A subset A of a metric space X is said to be **of first category** (or **meagre**) **in** X if it can be represented as a countable union of nowhere dense subsets of X. Otherwise, it is said to be **of second category** (or **nonmeagre**) **in** X. A set $B \subset X$ is said to be **residual in** X if ${}^cB$ is of first category in X.

To give an example of a set of first category arising naturally in a non-trivial context, we establish the following result.

Theorem 2.2.25 *Let X be a metric space, let (f_n) be a sequence of continuous real-valued functions on X which is pointwise convergent, and let the function $f : X \to \mathbf{R}$ be defined by*

$$ f(x) = \lim_{n \to \infty} f_n(x) \ (x \in X). $$

Then

$$ \mathscr{D} := \{x \in X : f \text{ is not continuous at } x\} $$

is of first category in X.

Proof Let ω be the oscillation function of f (see Definition 2.1.31). The identity

$$\mathscr{D} = \cup_{n=1}^{\infty}\{x \in X : \omega(x) \geq n^{-1}\}$$

exhibits $\mathscr{D}$ as a countable union of closed sets each of which will be shown to be of first category (in X). Plainly, a countable union of sets of first category is itself of first category. Thus $\mathscr{D}$ is of first category (in X).

Let $\varepsilon > 0$ and let

$$F = \{x \in X : \omega(x) \geq \varepsilon\}.$$

It is enough to establish that F is of first category. To do this, for $n \in \mathbf{N}$ let

$$E_n := \cap_{i,j \geq n}\{x \in X : \left|f_i(x) - f_j(x)\right| \leq \varepsilon/8\} :$$

each E_n is closed, $E_n \subset E_{n+1}$ and $X = \cup_{n=1}^{\infty} E_n$. Evidently

$$F = \cup_{n=1}^{\infty}(F \cap E_n),$$

and the matter of category is settled provided that, for each n, $\overset{o}{\overline{F \cap E_n}} = \emptyset$. To obtain a contradiction, suppose that for some n, $F \cap E_n$ is not nowhere dense. Then there exists an open set U such that

$$U \neq \emptyset, U \subset \overline{F \cap E_n} = F \cap E_n;$$

moreover, for each $x \in U$,

$$\left|f_i(x) - f_j(x)\right| \leq \varepsilon/8 \text{ if } i, j \geq n.$$

Setting $i = n$ and letting $j \to \infty$, it follows that

$$|f_n(x) - f(x)| \leq \varepsilon/8 \ (x \in U).$$

Let $y \in U$. Since f_n is continuous at y, there is a neighbourhood U_y of y such that $U_y \subset U$ and

$$|f_n(x) - f_n(y)| \leq \varepsilon/8 \ (x \in U_y).$$

It follows that

$$|f(x) - f_n(y)| \leq \varepsilon/4 \ (x \in U_y);$$

that

$$\left|f(x) - f(x')\right| \leq \varepsilon/2 \ (x, x' \in U_y);$$

and that

$$\omega(y) \leq \varepsilon/2.$$

But the last inequality, valid for all $y \in U$, implies that $U \cap F = \emptyset$, a conclusion incompatible with $U \neq \emptyset$, $U \subset F$. Thus, for all $n \in \mathbf{N}$, $F \cap E_n$ is nowhere dense, as required. □

Paraphrasing, Theorem 2.2.25 shows that, for an arbitrary metric space X, the set $\mathscr{D}$ of points of discontinuity of a function f generated as a pointwise limit of a sequence of continuous real-valued functions is of first category. Naturally, isolation of those metric spaces in which more can be said about $\mathscr{D}$ is of interest. If as an ideal one might wish to have $\mathscr{D} = \emptyset$, ${}^c\mathscr{D} = X$, then as a step towards this, generally, for complete metric spaces it turns out that $\overset{o}{\mathscr{D}} = \emptyset$, $\overline{{}^c\mathscr{D}} = X$. This is a consequence of the following theorem.

Theorem 2.2.26 *Let X be a metric space. If X has one of the following properties, then it has all of them.*

(i) *Every countable intersection of dense open subsets of X is dense in X.*
(ii) *The complement of every set of first category in X is dense in X.*
(iii) *Every set of first category in X has empty interior in X.*
(iv) *Every non-empty open set in X is of second category in X.*

A metric space with one, and hence all, of the above properties is said to be a ***Baire space****.*

Proof (i) $\Longrightarrow$ (ii) Let A be a set of first category in $X : A = \cup_{n=1}^{\infty} H_n$, where $\overset{o}{\overline{H_n}} = \emptyset$ $(n \in \mathbf{N})$. Let $B = \cup_{n=1}^{\infty} \overline{H_n}$. Then B is of first category in X and $A \subset B$. Now

$$ {}^cB = \cap_{n=1}^{\infty} {}^c(\overline{H_n}) \text{ and } \overline{{}^c(\overline{H_n})} = {}^c\left(\overset{o}{\overline{H_n}}\right) = X \ (n \in \mathbf{N}). $$

Hence, given that (i) holds and that ${}^cB \subset {}^cA$, it follows that $X = \overline{{}^cB} \subset \overline{{}^cA} \subset X$ and that ${}^cA$ is dense in X.

(ii) $\Longrightarrow$ (iii) Let A be of first category in X. Using (ii) we see that $\overline{{}^cA} = X$. Thus $\emptyset = {}^c(\overline{{}^cA}) = \overset{o}{A}$.

(iii) $\Longrightarrow$ (iv) To obtain a contradiction, suppose that U is a non-empty open subset of X which is of first category in X. Then, since (iii) holds, $\emptyset = \overset{o}{U} = U$.

(iv) $\Longrightarrow$ (i) Let $(\mathscr{O}_n)$ be a sequence of dense open subsets of X and let $E = \cap_{n=1}^{\infty} \mathscr{O}_n$. Then ${}^cE = \cup_{n=1}^{\infty} {}^c\mathscr{O}_n$ and

$$ \overset{o}{\overline{{}^c\mathscr{O}_n}} = \overset{o}{\widehat{{}^c\mathscr{O}_n}} = {}^c(\overline{\mathscr{O}_n}) = \emptyset \ (n \in \mathbf{N}). $$

Hence ${}^cE$ is of first category in X and consequently so is $\overset{o}{\widehat{{}^cE}}$. Since (iv) holds, it follows that $\emptyset = \overset{o}{\widehat{{}^cE}}$ and $X = \overline{E}$. □

In view of the above result it is obvious that our first version of Baire's theorem may now be recast in a final one as follows.

Theorem 2.2.27 *Every complete metric space is a Baire space.*

The reader should note that there exist incomplete metric spaces which are Baire spaces (see [2], Sect. 5, Exercise 14).

We conclude this section with the observation that, in the context of complete metric spaces, Baire's theorem immediately permits the following strengthened version of Theorem 2.2.25.

Theorem 2.2.28 *Let X be a complete metric space and let $f : X \to \mathbb{R}$ be the pointwise limit of a sequence of continuous real-valued functions on X. Then the set of points of continuity of f is residual and dense in X.*

Exercise 2.2.29

1. Let (X_1, d_1) and (X_2, d_2) be complete metric spaces, let $X = X_1 \times X_2$ and define $d : X \times X \to \mathbf{R}$ by
$$d(x, y) = \left\{d_1^2(x_1, y_1) + d_2^2(x_2, y_2)\right\}^{1/2} \quad (x = (x_1, x_2), y = (y_1, y_2) \in X).$$
Prove that (X, d) is complete.
2. Let (X, d) be a metric space and let F be a non-empty subset of X. Prove that
(i) if (X, d) is complete and F is closed relative to (X, d), then (F, d) is complete;
(ii) if (F, d) is complete, then F is closed relative to (X, d).
$\big[$By convention, (F, d) stands in place of $(F, d\,|_{F\times F})$.$\big]$
3. Let $I = [0, 1]$ and define $T : C(I) \to C(I)$ by
$$(Tf)(x) = x + \int_0^x (x - t)f(t)dt \;\; (x \in I, f \in C(I)).$$
Show that T is a contraction on $C(I)$ (assumed to be endowed with the uniform metric) and deduce that the only element f of $C(I)$ such that
$$f(x) = x + \int_0^x (x - t)f(t)dt \;\; (x \in I)$$
is the restriction to $[0, 1]$ of the hyperbolic sine function.
4. Use the contraction mapping theorem to show that for each $k \in (0, 1)$ the equation
$$f(x) = 1 + \int_0^x f(t^2)dt \;\; (0 \le x \le k)$$

has exactly one solution $f \in C([0, k])$. Hence show that this result is also true when $k = 1$.

5. (i) Give an example of a contraction mapping of an incomplete metric space into itself which has no fixed point.
   (ii) Give an example of a mapping T of a complete metric space (X, d) into itself with the property

$$d(Tx, Ty) < d(x, y) \text{ for all } x, y \in X, x \neq y,$$

   but which has no fixed point.
   (iii) Give an example of a mapping T of a complete metric space into itself such that T^m is a contraction mapping for some $m \in \mathbf{N}$, but T is not a contraction.
6. Let $X = \{x \in \mathbf{R} : 0 < x \leq 1\}$ and let d_1 and d_2 be metrics on X defined by

$$d_1(x, y) = |x - y|, d_2(x, y) = \left|\frac{1}{x} - \frac{1}{y}\right| \quad (x, y \in X).$$

   Prove that the two metric spaces (X, d_1) and (X, d_2) have the same convergent sequences, but that (X, d_2) is complete while (X, d_1) is not complete.
7. Let S be the set of all real sequences $x = (x_n)$ and let $d : S \times S \to \mathbf{R}$ be defined by

$$d(x, y) = \sum_{n=1}^{\infty} \frac{|x_n - y_n|}{2^n \left[1 + |x_n - y_n|\right]} \quad (x = (x_n), y = (y_n) \in S).$$

   Prove that (S, d) is a complete metric space.
8. Let (X, d) be a metric space.

   (i) Show that if (x_n) and (y_n) are Cauchy sequences in X, then $(d(x_n, y_n))$ is a Cauchy sequence in $\mathbf{R}$ and is therefore convergent.
   (ii) Let $\mathscr{X}$ be the set of all Cauchy sequences in X. Call elements (x_n), (y_n) of $\mathscr{X}$ equivalent, and write $(x_n) \sim (y_n)$, if $\lim_{n\to\infty} d(x_n, y_n) = 0$. Show that $\sim$ is an equivalence relation on $\mathscr{X}$.
   (iii) Let (x_n), (x_n'), (y_n) and $(y_n') \in \mathscr{X}$ and suppose that $(x_n) \sim (x_n')$ and $(y_n) \sim (y_n')$. Show that

$$\lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x_n', y_n').$$

   (iv) For $(x_n) \in \mathscr{X}$, let $[(x_n)]$ denote the equivalence class of which it is a member:

$$[(x_n)] = \{(y_n) \in \mathscr{X} : (y_n) \sim (x_n)\}.$$

   Let $\widehat{X}$ be the set of all equivalence classes and define $\widehat{d} : \widehat{X} \times \widehat{X} \to \mathbf{R}$ by

$$\widehat{d}\,([(x_n)], [(y_n)]) = \lim_{n\to\infty} d(x_n, y_n).$$

Show that $\widehat{d}$ is a metric on $\widehat{X}$ (it is well-defined by virtue of (iii)),

(v) For each $x \in X$, let $\phi(x) = [(x_n)]$, where $x_n = x$ for all $n \in \mathbf{N}$. Let $X_0 = \{\phi(x) : x \in X\}$. Show that, if X_0 is equipped with the metric inherited from $\widehat{X}$, then $x \mapsto \phi(x) : X \to X_0$ is an isometry.

(vi) Prove that X_0 is dense in $(\widehat{X}, \widehat{d})$, i.e., $\overline{X_0} = \widehat{X}$.

(vii) Prove that $(\widehat{X}, \widehat{d})$ is a complete metric space.

2.3 Compactness

We focus here on those metric spaces X with the following property: if a map $f : X \to \mathbf{R}$ is continuous then it is bounded, that is, its range, $f(X)$, is bounded. Spaces with this property are precisely those for which the sets $C(X)$ and $\mathscr{C}(X)$ coincide, a coincidence already noted in the case of each non-degenerate, closed, bounded interval in $\mathbf{R}$. In seeking to ensure the property three main strategies have emerged. These we now examine in turn.

Strategy I Let $f : X \to \mathbf{R}$ be continuous. Then each $x \in X$ has a neighbourhood U_x such that, for all $u \in U_x$,

$$|f(u)| < 1 + |f(x)| .$$

Evidently $X = \bigcup_{x \in X} U_x$. Hence if a **finite** set $\{x_1, x_2, \ldots, x_m\} \subset X$ exists such that $X = \bigcup_{k=1}^{m} U_{x_k}$, then $f(X)$ is bounded, since for all $u \in X$,

$$|f(u)| < 1 + \max_{1 \le k \le m} |f(x_k)| .$$

This observation motivates the next definition and establishes the theorem that follows.

Definition 2.3.1 A metric space X is said to be **compact** if every family $\mathscr{U}$ of open subsets of X such that $X = \cup \mathscr{U}$ contains a finite subfamily $\mathscr{V}$ such that $X = \cup \mathscr{V}$.

Theorem 2.3.2 *If X is a compact metric space, then every continuous map $f : X \to \mathbf{R}$ is bounded.*

By way of illustration, let $a, b \in \mathbf{R}$ and $a < b$. We claim that, viewed as a subspace of $\mathbf{R}$, the interval $[a, b]$ is compact. For if this were not so, then there would be a family $\mathscr{U}$ of open subsets of $[a, b]$, with union $[a, b]$, such that no finite collection of sets in $\mathscr{U}$ has union $[a, b]$. Bisect $[a, b]$: then at least one of the sub-intervals $[a, \frac{1}{2}(a + b)]$, $[\frac{1}{2}(a + b), b]$ is not contained in the union of any finite collection of members of $\mathscr{U}$. Repetition of this process gives a sequence of nested, closed sub-intervals of $[a, b]$, (I_n) say, with the length of I_n equal to $2^{-n}(b - a)$. By Cantor's intersection theorem (Theorem 2.2.3) these intervals I_n have intersection consisting of a single-point set in $[a, b]$, $\{x\}$ say. Obviously, there exists $U \in \mathscr{U}$ such that $x \in U$;

since U is open, $I_n \subset U$ for all large enough n. This contradicts the fact that no I_n is contained in the union of a finite number of members of $\mathscr{U}$, and our claim is justified.

Strategy II Suppose that X does not have the required property and that $f : X \to \mathbf{R}$ is a continuous but unbounded map. Then, for each $n \in \mathbf{N}$, there exists $x_n \in X$ such that $|f(x_n)| \geq n$. The sequence (x_n) does not have a convergent subsequence. To see this, suppose that (x_n) has a subsequence $(x_{m(n)})$ which converges to an element $x \in X$. Since f is continuous at x,

$$\left|f(x_{m(n)})\right| \to |f(x)|\,;$$

however, for all $n \in \mathbf{N}$, $\left|f(x_{m(n)})\right| \geq m(n) \geq n$, and so $\left|f(x_{m(n)})\right| \to \infty$.

We have shown that any metric space X without the required property has a sequence with no convergent subsequence. Put equivalently, if each sequence in X has a convergent subsequence, then each continuous, real-valued function on X is bounded. These matters are summarised below.

Definition 2.3.3 A metric space X is said to be **sequentially compact** if each sequence in X has a subsequence which converges to a point of X.

Theorem 2.3.4 *If X is a sequentially compact metric space, then every continuous map $f : X \to \mathbf{R}$ is bounded.*

That each closed, bounded interval in $\mathbf{R}$ is sequentially compact is immediate from the Bolzano-Weierstrass theorem.

We preface the final strategy with a definition.

Definition 2.3.5 A metric space X is said to be **totally bounded** if to each $\varepsilon > 0$ there corresponds a finite family $\mathscr{F}$ of subsets of X such that $X = \cup\mathscr{F}$ and, for each $F \in \mathscr{F}$, $\operatorname{diam} F < \varepsilon$.

Plainly, the interval $[0, 1]$, inheriting the usual metric from $\mathbf{R}$, is totally bounded. Indeed, every bounded subspace of $\mathbf{R}$ is totally bounded.

Strategy III Let X be complete and totally bounded. To obtain a contradiction, suppose that it carries a continuous but unbounded map $f : X \to \mathbf{R}$. Since X is totally bounded it is a union of finitely many closed sets each with diameter ≤ 1. (Observe that, if $A \subset X$, then $\operatorname{diam} A = \operatorname{diam} \overline{A}$.) The restriction of f to one of these, X_1 say, is unbounded. A further appeal to the total boundedness of X shows that it, and therefore X_1, is a union of finitely many sets each of which is closed in X and of diameter $\leq 1/2$. The restriction of f to one of these subsets of X_1, X_2 say, is unbounded. Proceeding in this way, the result is a sequence (X_n) of sets closed in X such that, for all $n \in \mathbf{N}$, (i) $X_{n+1} \subset X_n$, (ii) $\operatorname{diam} X_n \leq 1/n$, (iii) $f(X_n)$ is unbounded. By the Cantor intersection theorem, there exists $x \in X$ such that $\{x\} = \cap_{n=1}^{\infty} X_n$. Since f is continuous at x, there is a neighbourhood U_x of x on which f is bounded. But, for sufficiently large n, $X_n \subset U_x$ and therefore $f(X_n)$ is bounded. This contradicts (iii), and we have proved the following theorem.

Theorem 2.3.6 *If X is a complete and totally bounded metric space, then every continuous map $f : X \to \mathbf{R}$ is bounded.*

Conditions sufficient to ensure our property of interest are offered by each of Theorems 2.3.2, 2.3.4 and 2.3.6. Remarkably, they are also necessary conditions and hence equivalent. The position is formalised in our next result, the definition preceding which concerns terminology useful in its proof.

Definition 2.3.7 Let S be a subset of a set X and let $\mathscr{F}$ be a family of subsets of X such that $S \subset \cup\mathscr{F}$. Then $\mathscr{F}$ is called a **covering of** S : if $\mathscr{F}$ has only a finite number of members then it is called a **finite covering of** S. If X is a metric space and the members of $\mathscr{F}$ are open sets, $\mathscr{F}$ is called an **open covering of** S.

Theorem 2.3.8 *Let X be a metric space. The following are equivalent statements:*

(a) *X is compact.*
(b) *$C(X) = \mathscr{C}(X)$, i.e. each continuous map $f : X \to \mathbf{R}$ is bounded.*
(c) *X has the* ***Bolzano-Weierstrass property*** *: every infinite subset of X has a limit point in X.*
(d) *X is sequentially compact.*
(e) *X is complete and totally bounded.*

Proof (a)$\Longrightarrow$(b): this has already been established in Theorem 2.3.2.

(b)$\Longrightarrow$(c): let (b) hold and suppose that X has an infinite subset with no limit point in X. Then there is a sequence (x_n) of distinct points of X such that $S := \{x_n : n \in \mathbf{N}\}$ also has no limit point in X (the countable axiom of choice A.5.2 is used here). Thus, for each $n \in \mathbf{N}$, there exists $r_n \in (0, 1/n)$ such that $B(x_n, r_n) \cap S = \{x_n\}$. Letting d denote the metric on X, for each $n \in \mathbf{N}$ define $f_n : X \to \mathbf{R}$ by

$$f_n(x) = \max\left\{n\left(1 - 2r_n^{-1}d(x, x_n)\right), 0\right\}.$$

Evidently, each f_n is continuous.

Now, each element of X has a neighbourhood of itself restricted to which all save finitely many f_n are identically zero. To see this, let $x \in X$ and let $\rho > 0$ be such that $B(x, \rho) \cap S \subset \{x\}$. Further, let $N \in \mathbf{N}$ be such that $N > \rho^{-1}$ and, for all $n \geq N$, $x_n \neq x$. Then

$$B(x, \rho/2) \cap \bigcup_{n=N}^{\infty} B(x_n, r_n/2) = \emptyset;$$

for otherwise an $n \geq N$ would exist such that

$$\rho \leq d(x, x_n) < (\rho + r_n)/2 < (\rho + 1/n)/2 < \rho,$$

which is impossible. Hence, for all $n \geq N$ and all $y \in B(x, \rho/2)$, $f_n(y) = 0$.

Define $f : X \to \mathbf{R}$ by

$$f(x) = \sum_{n=1}^{\infty} f_n(x).$$

The existence for each $x \in X$ of an N and a ρ, both depending on x, such that

$$f(y) = \sum_{n=1}^{N} f_n(y) \quad (y \in B(x, \rho/2)),$$

shows that f is continuous: the fact that, for each $n \in \mathbf{N}, f(x_n) = n$, shows that it is unbounded. Such an f is incompatible with (b) and so X has the Bolzano-Weierstrass property.

(c)$\Longrightarrow$(d): Suppose that (c) holds and let (x_n) be a sequence in X. If there is a point $x \in X$ such that $x_n = x$ for infinitely many values of n, then evidently there is a subsequence of (x_n) which is constant (has all its terms equal to x) and converges to x. If no such x exists then $S := \{x_n : n \in \mathbf{N}\}$ is infinite and has a limit point $y \in X$. We now choose $m(1)$ to be the least positive integer n such that $0 < d(y, x_n) < 1$, and define inductively a subsequence $\left(x_{m(n)}\right)$ of (x_n) which converges to y. Suppose that $m(1) < m(2) < \ldots < m(n)$ have been chosen so that $0 < d(y, x_{m(j)}) < 1/j$ for $j = 1, 2, \ldots, n$. Choose $m(n+1)$ to be the least integer exceeding $m(n)$ such that $0 < d(y, x_{m(n+1)}) < 1/(n+1)$. This establishes (d).

(d)$\Longrightarrow$(e): Suppose that (d) holds, let $\varepsilon > 0$ and select $x_1 \in X$. Suppose that $x_1, \ldots, x_n$ have been chosen in X so that $d(x_i, x_j) \geq \varepsilon/3$ if $i \neq j$. If possible, choose $x_{n+1} \in X$ such that $d(x_i, x_{n+1}) \geq \varepsilon/3$ for all i, $1 \leq i \leq n$. This process must stop after a finite number of steps because of our assumption that (d) holds. It follows that $X = \bigcup_{j=1}^{N} B(x_j, \varepsilon/3)$ for some $N \in \mathbf{N}$. Since $B(x_j, \varepsilon/3)$ has diameter $< \varepsilon$, X is totally bounded.

To prove X complete, let (x_n) be a Cauchy sequence in X. By (d), there is a subsequence $(x_{k(n)})$ of (x_n) which converges to a point in X. Suppose $\lim_{n\to\infty} x_{k(n)} = x$ and let $\eta > 0$. There exists $N \in \mathbf{N}$ such that $d(x_m, x_n) < \eta/2$ and $d(x, x_{k(n)}) < \eta/2$ if $m, n \geq N$. Thus, for all $n \geq N$,

$$d(x, x_n) \leq d(x, x_{k(n)}) + d(x_{k(n)}, x_n) < \eta;$$

hence $\lim_{n\to\infty} x_n = x$; X is complete; and (e) holds.

(e)$\Longrightarrow$(a): Let $\mathscr{U}$ be an open covering of X and suppose that no finite subfamily of $\mathscr{U}$ is a covering of X. By (e), X is a union of finitely many closed sets each with diameter ≤ 1. One of these, say X_1, cannot be covered by finitely many members of $\mathscr{U}$. Repeat this argument with X_1 in place of X and continue indefinitely: there is a sequence (X_n) of closed sets such that, for all $n \in \mathbf{N}$, (i) $X_{n+1} \subset X_n$, (ii) diam $(X_n) < 1/n$, (iii) X_n is not covered by a finite subfamily of $\mathscr{U}$. By the Cantor intersection theorem there is a point $x \in X$ such that $\{x\} = \cap_{n=1}^{\infty} X_n$. Hence $x \in U$ for some $U \in \mathscr{U}$. By (ii), $X_n \subset U$ for all large enough n and this contradicts (iii). Hence X is compact and the proof of the theorem is complete. $\square$

Corollary 2.3.9 *If X is compact then $C(X)$, equipped with the uniform metric*

$$d_\infty(f, g) = \sup_{x \in X} |f(x) - g(x)|,$$

is complete.

Proof The result is immediate from Theorems 2.2.6 and 2.3.8. □

A standard method generates new compact spaces from old. Proceeding through a reformulation of total boundedness we show that a finite Cartesian product of compact spaces equipped with the standard metric is compact.

Lemma 2.3.10 *Let (X, d) be a metric space. The following three statements are equivalent.*

(a) *X is totally bounded.*
(b) *Given any $\varepsilon > 0$, there exists a finite set $F \subset X$ such that, for each $x \in X$,*

$$d(x, F) := \inf \{d(x, f) : f \in F\} < \varepsilon.$$

(c) *Given any $\varepsilon > 0$, there exists a finite set $F \subset X$ such that $X \subset \bigcup_{f \in F} B(f, \varepsilon)$.*

Proof Suppose that (a) holds, that $\varepsilon > 0$ and that $x \in X$. There exist finitely many non-empty sets $A_1, A_2, \ldots, A_k \subset X$, whose union covers X, and each of which has diameter less than ε. Select $a_j \in A_j$ for $1 \le j \le k$; put $F = \{a_1, a_2, \ldots, a_k\}$. Then, for some j, $x \in A_j$ and $d(x, a_j) < \varepsilon$. Hence $d(x, F) < \varepsilon$ and (b) holds.

Next, suppose that (b) holds and that $\varepsilon > 0$. Let F be a finite set in X such that, for each $x \in X$, $d(x, F) < \varepsilon$. Plainly $X = \cup_{f \in F} B(f, \varepsilon)$, since, for each $x \in X$ there is an $f \in F$ with $d(x, f) < \varepsilon$. Thus (b) implies (c).

Finally, suppose that (c) holds and that $\varepsilon > 0$. Evidently $X = \cup_{f \in F} B(f, \varepsilon/3)$ for some finite set $F \subset X$, and so X is covered by finitely many sets each with diameter $< \varepsilon$. Hence (a) holds. □

Corollary 2.3.11 *If X is totally bounded, then it is bounded.*

Proof Let X be totally bounded. There exists a finite set $F \subset X$ such that for all $x \in X$, $d(x, F) < 1$. Hence $\operatorname{diam} X \le 2 + \operatorname{diam} F$. □

Theorem 2.3.12 *Let $(X_1, d_1), \ldots, (X_n, d_n)$ be compact metric spaces. Let $X = \prod_{k=1}^n X_k$ and let d be defined for all $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in X$ by*

$$d(x, y) = \left\{ \sum_{k=1}^n d_k^2(x_k, y_k) \right\}^{1/2}.$$

Then (X, d) is a compact metric space.

Proof The obvious extension of Exercise 2.2.29 /1 shows that (X, d) is complete. It remains to prove that it is totally bounded.

Let $\varepsilon > 0$. Use of Lemma 2.3.10 shows that for each $k \in \{1, \ldots, n\}$, there is a finite set $F_k \subset X_k$ such that, for all $u \in X_k$, $d_k(u, F_k) < \varepsilon/\sqrt{n}$. Let $F = \prod_{k=1}^{n} F_k$: F is finite. Let $x = (x_1, \ldots, x_n) \in X$. For each k, there exists $f_k \in F_k$ such that $d_k(x_k, f_k) < \varepsilon/\sqrt{n}$. Let $f = (f_1, \ldots, f_n)$: $f \in F$ and $d(x, f) < \varepsilon$. Hence (X, d) is totally bounded. □

Metric spaces are often encountered as subspaces of others. Language for the situation in which the embedded space is compact is introduced in the next definition. The two lemmas which follow it describe attributes of a compact subspace relative to its host.

Definition 2.3.13 A subset of a metric space X is said to be a **compact set** in X if either it is empty or it is compact as a subspace of X. The obvious substitutions respectively define (a) a **sequentially compact set** in X, (b) a **totally bounded set** in X.

Lemma 2.3.14 *Let E be a subset of a metric space X. Then E is a compact set in X if, and only if, every covering of E by sets open in X contains a finite covering of E.*

Proof If $E = \emptyset$ the result is obvious.

Let E be a non-empty compact set in X and let $\mathscr{U}$ be a covering of E by sets open in X. The family $\widetilde{\mathscr{U}} = \{E \cap U : U \in \mathscr{U}\}$ is a covering of E by sets open in the subspace E and, therefore, there exist sets $U_1, \ldots, U_n \in \mathscr{U}$ such that $E = \cup_{j=1}^{n} E \cap U_j \subset \cup_{j=1}^{n} U_j$. Thus $\mathscr{U}$ contains a finite covering of E.

Conversely, let E be non-empty and let $\mathscr{V}$ be a covering of E by sets open in the subspace E. Then there exists a covering $\mathscr{U}$ of E, whose elements are open in the metric space X, such that $\mathscr{V} = \{E \cap U : U \in \mathscr{U}\}$. Since there are sets $U_1, \ldots, U_n \in \mathscr{U}$ such that $E \subset \cup_{j=1}^{n} U_j$, $\mathscr{V}$ contains a finite covering of E. Hence E is a compact set in X. □

Lemma 2.3.15 *Let X be a metric space.*

(i) *If E is a compact set in X then it is closed and bounded in X.*
(ii) *If X is a compact space and E is a closed set in X, then E is a compact set in X.*

Proof (i) If $E = \emptyset$ then the matter is clear.

Suppose $E \neq \emptyset$. The subspace E is complete and totally bounded. Hence the set E is closed in the space X (Exercise 2.2.29/2(ii)) and bounded in the space X, since it is bounded in the subspace E (Corollary 2.3.11).

(ii) Suppose $E \neq \emptyset$; otherwise the result holds trivially. Since E is closed in X, E is a complete subspace of X (Exercise 2.2.29/2(i)). Since X is a totally bounded space, so also is the subspace E. Hence E is a compact set in X. □

For general spaces X, the converse of Lemma 2.3.15 (i) is false. However, in the important special case when $X = \mathbf{R}^n$, it is true.

Theorem 2.3.16 (Heine-Borel) *Let K be a subset of $\mathbf{R}^n$. Then K is a compact set in $\mathbf{R}^n$ if, and only if, it is closed and bounded.*

Proof If K is compact in $\mathbf{R}^n$ then by Lemma 2.3.15 (i), it is a closed and bounded set in $\mathbf{R}^n$.

Conversely, suppose that K is a non-empty, closed and bounded set in $\mathbf{R}^n$; if $K = \emptyset$ then the result holds trivially. Observe that K is contained in a cube $I^n \subset \mathbf{R}^n$, where I is a closed and bounded interval in $\mathbf{R}$. Since K is closed in $\mathbf{R}^n$ it is closed in the subspace I^n. Hence, by Lemma 2.3.15 (ii), if I^n is compact then K is a compact set in I^n and also in $\mathbf{R}^n$. It remains to prove that I^n is compact.

Note that I regarded as a subspace of $\mathbf{R}$ is complete and totally bounded and therefore compact. By Theorem 2.3.12, I^n is compact, as required. □

The example to follow reinforces the fact that the converse of Lemma 2.3.15 (i) is false. It is sited in ℓ_2, and is complemented by a characterisation of the compact sets therein.

Example 2.3.17 From Examples 2.1.2 (vi) and 2.2.2 (vi),

$$\ell_2 = \left\{ x = (x_n)_{n\in\mathbf{N}} : x_n \in \mathbf{R} \text{ for all } n \in \mathbf{N}, \sum_{n=1}^{\infty} x_n^2 < \infty \right\}$$

is a complete metric space when equipped with the metric

$$d(x, y) = \left\{ \sum_{n=1}^{\infty} (x_n - y_n)^2 \right\}^{1/2},$$

where $x = (x_n), y = (y_n) \in \ell_2$.

Let $K = \{x \in \ell_2 : d(0, x) \leq 1\}$: K is the closed ball with centre 0 and radius 1. Although closed and bounded, K is not compact. For let $e^n = (\delta_j^n)_{j\in\mathbf{N}} \in \ell_2$, where δ_j^n is the Kronecker delta, equal to 1 when $j = n$ and zero otherwise. As the sequence (e^n) in ℓ_2 is such that $d(e^m, e^n) = \sqrt{2}$ if $m \neq n$, it has no convergent subsequence.

Theorem 2.3.18 *Let A be a non-empty subset of ℓ_2. Then A is compact if, and only if, it is closed, bounded and such that*

$$\sup_{x=(x_k)\in A} \sum_{k=n}^{\infty} x_k^2 \to 0 \text{ as } n \to \infty. \qquad (2.3.1)$$

Proof Let A be compact. By Lemma 2.3.15 (i), A is closed and bounded. Let $\varepsilon > 0$. There exists a finite set $F = \{a^1, \ldots, a^p\} \subset A$ such that, for all $x \in A$, $d(x, F) < \varepsilon/2$. Choose $m \in \mathbf{N}$ such that

$$\max_{1\leq q\leq p} \left\{ \sum_{k=m}^{\infty} \left|a_k^q\right|^2 \right\}^{1/2} < \varepsilon/2.$$

Let $x = (x_k) \in A$. There exists r, $1 \le r \le p$, such that $d(x, a^r) < \varepsilon/2$. Thus, by the Minkowski inequality,

$$\left\{\sum_{k=m}^{\infty} |x_k|^2\right\}^{1/2} \le \left\{\sum_{k=m}^{\infty} \left|x_k - a_k^r\right|^2\right\}^{1/2} + \left\{\sum_{k=m}^{\infty} \left|a_k^r\right|^2\right\}^{1/2}$$

$$\le d(x, a^r) + \left\{\sum_{k=m}^{\infty} \left|a_k^r\right|^2\right\}^{1/2} < \varepsilon,$$

and so

$$\sup_{x=(x_k)\in A} \left\{\sum_{k=m}^{\infty} x_k^2\right\}^{1/2} \le \varepsilon.$$

Conversely, suppose that A is closed and bounded and has the property (2.3.1). Then A is closed in the complete space ℓ_2 and hence is a complete subspace of ℓ_2. It remains to show that it is totally bounded. Given $\varepsilon > 0$, choose $n \in \mathbf{N}$ such that

$$\sup_{x=(x_k)\in A} \left\{\sum_{k=n+1}^{\infty} x_k^2\right\}^{1/2} < \varepsilon/2.$$

Since A is bounded, there exists a real number Λ such that, for all $x \in A$, $d(0, x) \le \Lambda$. Let $\lambda = \left\{\sum_{k=1}^{\infty} k^{-2}\right\}^{1/2}$ and choose $m \in \mathbf{N}$ so that $m\varepsilon > 4\lambda\Lambda$. For each j, $1 \le j \le n$, let

$$F_j = \left\{\left(-1 + \frac{2r}{jm}\right)\Lambda : r = 0, 1, \ldots, jm\right\}.$$

Put

$$F = \{x = (x_k) \in \ell_2 : x_k \in F_k \text{ if } 1 \le k \le n, \text{ and } x_k = 0 \text{ if } k > n\} :$$

F is a finite set in ℓ_2. Let $x = (x_k) \in A$. There exists $f = (f_k) \in F$ such that, if $1 \le k \le n$, then $|x_k - f_k| \le k^{-1}(2\Lambda/m)$. Hence

$$d^2(x, f) = \sum_{k=1}^{n} |x_k - f_k|^2 + \sum_{k=n+1}^{\infty} |x_k|^2$$

$$< (2\lambda\Lambda/m)^2 + (\varepsilon/2)^2 < \varepsilon^2.$$

As it is covered by finitely many balls of radius ε, it follows that A is totally bounded. □

Corollary 2.3.19 *The Hilbert cube*

$$\mathscr{H} = \left\{ x = (x_k) \in \ell_2 : \textit{for each}\, k \in \mathbf{N}, |x_k| \le k^{-1} \right\}$$

is compact in ℓ_2.

Proof It is routine to check that $\mathscr{H}$ is closed and bounded in ℓ_2 and that (2.3.1) holds. □

Pursuing the characterisation of compact sets in special spaces a little further, we consider next the position in spaces kindred to $C[0, 1]$. The best known characterisation in such spaces is the Arzelà-Ascoli theorem, which involves the concept of equicontinuity explained below.

Definition 2.3.20 Let (X, d) be a metric space. A set $\mathscr{F} \subset C(X)$ is said to be **equicontinuous at a point** $x \in X$ if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $y \in X$ with $d(x, y) < \delta$, we have $\sup_{f \in \mathscr{F}} |f(y) - f(x)| < \varepsilon$; if $\mathscr{F}$ is equicontinuous at every point of X we say that $\mathscr{F}$ is **equicontinuous on** X.

Lemma 2.3.21 *Let* (X, d) *be a compact metric space and let* $\mathscr{F} \subset C(X)$. *Then* $\mathscr{F}$ *is equicontinuous on* X *if, and only if, it is* ***uniformly equicontinuous on*** X *in the sense that given any* $\varepsilon > 0$, *there exists* $\delta > 0$ *such that for all* $x, y \in X$ *with* $d(x, y) < \delta$, *we have* $\sup_{f \in \mathscr{F}} |f(y) - f(x)| < \varepsilon$.

Proof Suppose that $\mathscr{F}$ is equicontinuous on X and let $\varepsilon > 0$. Then given any $x \in X$, there exists $\delta_x > 0$ such that $\sup_{f \in \mathscr{F}} |f(y) - f(x)| < \varepsilon/3$ if $d(x, y) < \delta_x$. By the compactness of X, there exist $x_1, \ldots, x_n \in X$ such that $X = \cup_{j=1}^n B(x_j, \delta_{x_j}/2)$; let $\delta = \frac{1}{2} \min \left\{ \delta_{x_1}, \ldots, \delta_{x_n} \right\}$. Let $x, y \in X$, $d(x, y) < \delta$ and $f \in \mathscr{F}$. For some j, $x \in B(x_j, \delta_{x_j}/2)$. Thus x and y belong to $B(x_j, \delta_{x_j})$ and

$$|f(x) - f(y)| \le \left| f(x) - f(x_j) \right| + \left| f(x_j) - f(y) \right| < 2\varepsilon/3.$$

It follows that $\sup_{f \in \mathscr{F}} |f(y) - f(x)| < \varepsilon$ whenever $d(x, y) < \delta$: $\mathscr{F}$ is uniformly equicontinuous on X.

The converse is obvious. □

Theorem 2.3.22 (Arzelà-Ascoli) *Let* (X, d) *be a compact metric space and let* $\mathscr{K} \subset C(X)$. *Then* $\mathscr{K}$ *is compact if, and only if, it is closed, bounded and equicontinuous on* X.

Proof Suppose $\mathscr{K}$ is compact. Then it is certainly closed and bounded. To establish equicontinuity, let $\varepsilon > 0$ and let $f_1, \ldots, f_n \in \mathscr{K}$ be such that $\mathscr{K} \subset \cup_{k=1}^n B_{C(X)}(f_k, \varepsilon/3)$. Since each f_k is uniformly continuous on X, there exists $\delta > 0$ such that $|f_k(x) - f_k(y)| < \varepsilon/3$ if $d(x, y) < \delta$ and $k \in \{1, 2, \ldots, n\}$. Now let $f \in \mathscr{K}$ and $d(x, y) < \delta$: then $f \in B_{C(X)}(f_k, \varepsilon/3)$ for some k and

$$|f(x) - f(y)| \le |f(x) - f_k(x)| + |f_k(x) - f_k(y)| + |f_k(y) - f(y)| < \varepsilon;$$

that is, $\mathscr{K}$ is equicontinuous on X.

Conversely, suppose that $\mathscr{K}$ is equicontinuous on X, closed and bounded. It is enough to prove that $\mathscr{K}$ is totally bounded. To do so, let $\varepsilon > 0$. Since X is compact, $\mathscr{K}$ is uniformly equicontinuous on X and thus there is a $\delta > 0$ such that, whenever $x, y \in X$ and $d(x, y) < \delta$,

$$\sup_{f \in \mathscr{K}} |f(x) - f(y)| < \varepsilon/4;$$

also, there exist $x_1, \ldots, x_n \in X$ such that

$$X = \cup_{i=1}^{n} U_i,$$

where $U_i = B_X(x_i, \delta) \ \ (1 \leq i \leq n)$. Moreover, each $x \in X$ is such that, for some $i \in \{1, 2, \ldots, n\}$, $x \in U_i$ and

$$\sup_{f \in \mathscr{K}} |f(x) - f(x_i)| < \varepsilon/4.$$

Let $\Lambda = \sup_{f \in \mathscr{K}} d_\infty(f, 0)$, $I = \{\lambda \in \mathbf{R} : |\lambda| \leq \Lambda\}$ and define $\theta : \mathscr{K} \to I^n \subset \mathbf{R}^n$ by

$$\theta(f) = (f(x_1), \ldots, f(x_n)).$$

Since I^n is totally bounded, so also is $\theta(\mathscr{K})$. Thus there exist $f_1, \ldots, f_m \in \mathscr{K}$ such that

$$\theta(\mathscr{K}) \subset \cup_{j=1}^{m} B_{\mathbf{R}^n}(\theta(f_j), \varepsilon/4)$$

and, to conclude the proof, we show that

$$\mathscr{K} \subset \cup_{j=1}^{m} B_{C(X)}(f_j, \varepsilon).$$

Let $f \in \mathscr{K}$. For some $j \in \{1, 2, \ldots, m\}$,

$$d_{\mathbf{R}^n}\left(\theta(f), \theta(f_j)\right) = \left\{\sum_{i=1}^{n} \left|f(x_i) - f_j(x_i)\right|^2\right\}^{1/2} < \varepsilon/4,$$

and therefore

$$\max_{1 \leq i \leq n} \left|f(x_i) - f_j(x_i)\right| < \varepsilon/4.$$

Let $x \in X$. Then, for some $i \in \{1, 2, \ldots, n\}$, $x \in U_i$ and

$$\max\left\{\left|f(x) - f(x_i)\right|, \left|f_j(x) - f_j(x_i)\right|\right\} < \varepsilon/4.$$

Hence

$$\left|f(x) - f_j(x)\right| \leq \left|f(x) - f(x_i)\right| + \left|f(x_i) - f_j(x_i)\right| + \left|f_j(x_i) - f_j(x)\right| < 3\varepsilon/4.$$

It follows that $d_\infty(f, f_j) < \varepsilon$ and $f \in B_{C(X)}(f_j, \varepsilon)$. The proof is complete. □

Corollary 2.3.23 *Let X be a compact metric space and let $\mathscr{K} \subset C(X)$. Then $\mathscr{K}$ is relatively compact (that is, $\overline{\mathscr{K}}$ is compact) if, and only if, it is bounded and equicontinuous.*

Proof It is easy to prove that if $\mathscr{K}$ is equicontinuous on X, then so is $\overline{\mathscr{K}}$. The rest is obvious. □

We next turn to continuous maps on compact spaces.

Theorem 2.3.24 *Let X_1 and X_2 be metric spaces and let $f : X_1 \to X_2$ be continuous.*

(i) *If E is a compact set in X_1, then $f(E)$ is a compact set in X_2.*
(ii) *If X_1 is compact and f is bijective, then f is a homeomorphism.*

Proof

(i) Suppose $E \neq \emptyset$; otherwise the result holds trivially. Let $\mathscr{V}$ be an open covering of $f(E)$. Then $\mathscr{U} = \{f^{-1}(V) : V \in \mathscr{V}\}$ is an open covering of E. Since E is compact in X_1, by Lemma 2.3.14, $\mathscr{U}$ contains a finite covering, $\{f^{-1}(V_1), \ldots, f^{-1}(V_n)\}$ say, of E. But this implies that $\{V_1, \ldots, V_n\} \subset \mathscr{V}$ is a finite covering of $f(E)$. Using Lemma 2.3.14 again, we see that $f(E)$ is a compact set in X_2.
(ii) Let U be an open set in X_1. Since $X_1 \backslash U$ is closed and therefore compact in X_1 (Lemma 2.3.15 (ii)), by the first part of the theorem, $f(X_1 \backslash U)$ is compact and therefore closed in X_2 (Lemma 2.3.15 (i)). Further, since $f(X_1 \backslash U) = X_2 \backslash f(U)$, it follows that $f(U)$ is open in X_2. Now, appeal to Remark 2.1.37 (ii) completes the proof. □

Corollary 2.3.25 *Let X be a metric space, let $f : X \to \mathbf{R}$ be continuous, and let E be a non-empty compact set in X. Then $f(E)$ is bounded and both* $\inf f(E)$ *and* $\sup f(E)$ *belong to $f(E)$. In particular, there exist points u and v in E such that*

$$f(u) = \inf f(E) \text{ and } f(v) = \sup f(E).$$

Proof By Theorem 2.3.24, $f(E)$ is compact in $\mathbf{R}$; by Theorem 2.3.16, it is closed and bounded. The result now follows easily. □

The novelty of Corollary 2.3.25 is in the attainment of bounds. It is utilised in proving the next result, about the distance of a point from a set and the distance between two sets.

Theorem 2.3.26 *Let A and B be non-empty subsets of a metric space (X, d), and let $d(x, A)$, $d(A, B)$ be defined as in Lemma 2.1.40. Then*

(i) *if A is compact, B is closed and $A \cap B = \emptyset$, then there exists $a \in A$ such that $d(a, B) = d(A, B) > 0$;*
(ii) *if $X = \mathbf{R}^n$, d is the Euclidean metric on $\mathbf{R}^n$, A is compact, B is closed and $A \cap B = \emptyset$, then there exist $a \in A$ and $b \in B$ such that $d(a, b) = d(A, B)$.*

Proof

(i) By Lemma 2.1.40 (iii), the map $x \longmapsto d(x, B) : X \to \mathbf{R}$ is continuous. Since A is compact, by Corollary 2.3.25 there exists $a \in A$ such that

$$d(a, B) = \inf \{d(x, B) : x \in A\}.$$

Using Lemma 2.1.40 (i), we therefore see that $d(a, B) = d(A, B)$. Now, if $d(a, B) = 0$ then $a \in \overline{B}$ (by Lemma 2.1.40 (ii)). Given that B is closed, it would follow that $a \in A \cap B$, which contradicts $A \cap B = \emptyset$.
(ii) By (i), there exists $a \in A$ such that $d(a, B) = d(A, B)$. Choose any $\widetilde{b} \in B$. Let $\widetilde{B} = \left\{y \in B : d(a, y) \leq d(a, \widetilde{b})\right\}$; note that $d(a, \widetilde{B}) = d(a, B)$. Since $\widetilde{B}$ is closed and bounded, by Theorem 2.3.16 it is compact. Thus, using (i), there exists $b \in \widetilde{B} \subset B$ such that $d(a, b) = d(b, \{a\}) = d(\widetilde{B}, \{a\}) = d(a, \widetilde{B})$ $= d(a, B)$. □

Note that the conclusion of (i) is false if the set A is merely required to be closed, rather than compact. To illustrate this take $X = \mathbf{R}$, with the usual metric, let $A = \mathbf{N}$, $B = \{n - 1/n : n \in \mathbf{N}\}$. Plainly A and B are closed, and $A \cap B = \emptyset$; but, for all $n \in \mathbf{N}$, $d(A, B) \leq d(n - 1/n, n) = 1/n \to 0$ as $n \to \infty$, and hence $d(A, B) = 0$.

Corollary 2.3.25 has many uses, and it is worthwhile to note that key aspects of it apply to functions with properties similar to continuity but of weaker regularity. Looking back at Definition 2.1.27 it is clear that for a real-valued function f on a metric space X to be continuous at $x \in X$, it is necessary and sufficient that, given any $\varepsilon > 0$, there exists a neighbourhood U of x such that
(i) $f(x) - \varepsilon < f(u)$ whenever $u \in U$,
and
(ii) $f(u) < f(x) + \varepsilon$ whenever $u \in U$.
Taken separately, conditions (i) and (ii) define classes of functions of importance in their own right.

Definition 2.3.27 Let (X, d) be a metric space, $x \in X$ and f be a real-valued function on X. Then f is said to be **lower semi-continuous at** x if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$f(x) - \varepsilon < f(y) \text{ if } d(x, y) < \delta.$$

The function f is said to be **lower semi-continuous on** X if it is lower semi-continuous at each point of X. Similarly, f is called **upper semi-continuous at** x if, given any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$f(y) < f(x) + \varepsilon \text{ if } d(x, y) < \delta;$$

upper semi-continuity on X is defined in the obvious way.

Plainly, given $x \in X$, a necessary and sufficient condition for f to be continuous at x is that it should be both lower semi-continuous at x and upper semi-continuous at x. Note that if f is lower semi-continuous at x, then $-f$ is upper semi-continuous at x.

Example 2.3.28

(1) Let X be a metric space and $x \in X$. A function $f : X \to \mathbf{R}$ is said to have a **relative minimum** (respectively, **relative maximum**) at x if there exists a neighbourhood U of x such that $f(x) \leq f(u)$ (respectively, $f(x) \geq f(u)$) whenever $u \in U$. It is clear that if f has a relative minimum (maximum) at x then it is lower (upper) semi-continuous at x.

(2) Let X be a metric space and $A \subset X$. Then A is open in X if, and only if, the characteristic function χ_A of A is lower semi-continuous on X. To see this, suppose that A is open. Then χ_A has a relative minimum at each point of X and so is lower semi-continuous on X. Conversely, let χ_A be lower semi-continuous on X. Omitting the trivial case of $A = \emptyset$, let $a \in A$. Then there is a neighbourhood V of a such that

$$\frac{1}{2} = \chi_A(a) - \frac{1}{2} < \chi_A(x) \text{ if } x \in V.$$

But this shows that $V \subset A$, that $A \subset \overset{o}{A}$ and that A is open.

(3) Let $f : [0, 1] \to \mathbf{R}$ be defined by $f(t) = 0$ if t is irrational, $f(t) = 1/q$ if $t = p/q$, where p and q are integers with no common factor greater than 1, and $q > 0$. Then f is upper semi-continuous on $[0, 1]$. Note that f is continuous and therefore upper semi-continuous at all irrational points of $[0, 1]$ (refer to Exercise 1.3.10/6). Further, it has a relative maximum at every rational point of $[0, 1]$.

Lemma 2.3.29 *Let X be a compact metric space and $f : X \to \mathbf{R}$ be lower semi-continuous on X. Then f has an **absolute minimum** on X : there exists $u \in X$ such that, for all $x \in X$, $f(u) \leq f(x)$.*

Proof We begin by showing that f is bounded below. Suppose otherwise. Then for each $n \in \mathbf{N}$, there exists $x_n \in X$ such that $f(x_n) \leq -n$. Hence

$$\lim_{n\to\infty} f(x_n) = -\infty. \tag{2.3.2}$$

Since X is compact, the sequence (x_n) has a convergent subsequence, $(x_{m(n)})$ say. Suppose $\lim_{n\to\infty} x_{m(n)} = x$. As f is lower semi-continuous at x, a neighbourhood V of x exists such that

$$f(x) - 1 < f(v) \ \ (v \in V).$$

Hence, for sufficiently large n,

$$f(x) - 1 < f(x_{m(n)}),$$

a conclusion incompatible with (2.3.2).

Now let

$$K = \inf\{f(x) : x \in X\};$$

it remains to show that $K \in f(X)$. For each $n \in \mathbf{N}$, there exists $u_n \in X$ such that $f(u_n) < K + n^{-1}$. Since X is compact, the sequence (u_n) has a convergent subsequence, $(u_{k(n)})$ say. Suppose $u_{k(n)} \to u$. Clearly

$$K \le f(u_{k(n)}) < K + k(n)^{-1} \ (n \in \mathbf{N})$$

and so

$$K = \lim_{n\to\infty} f(u_{k(n)}).$$

Further, as f is lower semi-continuous at u, for each $\varepsilon > 0$ there exists a neighbourhood U of u such that

$$f(u) - \varepsilon < f(x) \text{ whenever } x \in U.$$

Thus for all $\varepsilon > 0$,

$$f(u) - \varepsilon \le \lim_{n\to\infty} f(u_{k(n)}) = K,$$

and so $f(u) \le K$. But by definition of $K, f(u) \ge K$. Thus $K = f(u) \in f(X)$. □

Theorem 2.3.30 *Let* (X_1, d_1), (X_2, d_2) *be metric spaces, let* (X_1, d_1) *be compact and let* $f : X_1 \to X_2$ *be continuous. Then* f *is uniformly continuous.*

Proof When (X_2, d_2) is $\mathbf{R}$ equipped with the standard metric, this result follows from Lemma 2.3.21. To illustrate the ways in which compactness can be used we give here a proof by contradiction that is not so readily available for Lemma 2.3.21. Suppose f is not uniformly continuous. Then there exists $\varepsilon > 0$ such that, given any $\delta > 0$, there exist $x, y \in X_1$ with

$$d_1(x, y) < \delta \text{ and } d_2(f(x), f(y)) \ge \varepsilon.$$

Hence, for all $n \in \mathbf{N}$ there exist $x_n, y_n \in X_1$ such that

$$d_1(x_n, y_n) < 1/n \text{ and } d_2(f(x_n), f(y_n)) \ge \varepsilon.$$

Since X_1 is compact, there exist $x \in X_1$ and a subsequence $(x_{m(n)})$ of (x_n) such that $d_1(x_{m(n)}, x) \to 0$. As $d_1(x_{m(n)}, y_{m(n)}) < 1/m(n) \leq 1/n \to 0$, it follows that $d_1(y_{m(n)}, x) \to 0$. But given that f is continuous,

$$\varepsilon \leq d_2(f(x_{m(n)}), f(y_{m(n)})) \leq d_2(f(x_{m(n)}), f(x)) + d_2(f(x), f(y_{m(n)})) \to 0,$$

a contradiction. Hence f is uniformly continuous. □

Corollary 2.3.31 *Let $f : \mathbf{R}^2 \to \mathbf{R}$ be continuous and have a continuous first partial derivative $\partial_2 f$ with respect to the second coordinate. For each $t \in \mathbf{R}$ put*

$$F(t) = \int_0^1 f(s, t)ds.$$

Then $F : \mathbf{R} \to \mathbf{R}$ is differentiable and, for all $t \in \mathbf{R}$,

$$F'(t) = \int_0^1 \partial_2 f(s, t)ds.$$

Proof Let $t_0, t \in \mathbf{R}$, $t \neq t_0$. Then

$$\left| \frac{F(t) - F(t_0)}{t - t_0} - \int_0^1 \partial_2 f(s, t_0)ds \right| \leq \int_0^1 \left| \frac{f(s, t) - f(s, t_0)}{t - t_0} - \partial_2 f(s, t_0) \right| ds.$$

Let $\varepsilon > 0$. By Theorem 2.3.16, $[0, 1] \times [t_0 - 1, t_0 + 1]$ is a compact subset of $\mathbf{R}^2$; thus the continuous map $\partial_2 f$ is uniformly continuous on this rectangle, by Theorem 2.3.30. Hence there exists $\delta \in (0, 1)$ such that, for all $s \in [0, 1]$ and all $v \in [t_0 - \delta, t_0 + \delta]$,

$$|\partial_2 f(s, v) - \partial_2 f(s, t_0)| < \varepsilon.$$

It follows that, if $s \in [0, 1]$ and $0 < |t - t_0| < \delta$, then (by the mean-value theorem) for some v strictly between t and t_0,

$$\left| \frac{f(s, t) - f(s, t_0)}{t - t_0} - \partial_2 f(s, t_0) \right| = |\partial_2 f(s, v) - \partial_2 f(s, t_0)| < \varepsilon.$$

Thus

$$\left| \frac{F(t) - F(t_0)}{t - t_0} - \int_0^1 \partial_2 f(s, t_0)ds \right| < \varepsilon$$

if $0 < |t - t_0| < \delta$, and the result follows. □

This Corollary is particularly useful. To illustrate, we give the following example.

Example 2.3.32 For each $t \in \mathbf{R}$, put $g(t) = \left(\int_0^t e^{-s^2} ds \right)^2$. By the fundamental theorem of integral calculus,

$$g'(t) = 2e^{-t^2}\int_0^t e^{-s^2}ds.$$

Put $h(t) = \int_0^1 (1+s^2)^{-1}e^{-t^2(1+s^2)}ds \;\; (t \in \mathbf{R})$. By Corollary 2.3.31,

$$h'(t) = -2t\int_0^1 e^{-t^2(1+s^2)}ds = -2te^{-t^2}\int_0^1 e^{-t^2s^2}ds.$$

If $t \neq 0$, the substitution $u = st$ reduces this to

$$h'(t) = -2e^{-t^2}\int_0^t e^{-u^2}du\,(t \neq 0);$$

plainly $h'(t) = 0$ if $t = 0$. Thus, for all $t \in \mathbf{R}$, $g'(t) + h'(t) = 0$ and so

$$g(t) + h(t) = \text{ constant } = g(0) + h(0) = \int_0^1 (1+s^2)^{-1}ds = \pi/4.$$

Hence

$$\left|\frac{\pi}{4} - g(t)\right| = h(t) = e^{-t^2}\int_0^1 (1+s^2)^{-1}e^{-t^2s^2}ds \leq e^{-t^2} \to 0$$

as $t \to \infty$. It follows that

$$\lim_{t\to\infty}\left(\int_0^t e^{-s^2}ds\right)^2 = \frac{\pi}{4},$$

which gives the famous result that

$$\int_0^\infty e^{-s^2}ds = \frac{1}{2}\sqrt{\pi}.$$

We conclude this section by giving two applications of the ideas of completeness and compactness.

2.3.1 Application 1

This is to the theory of ordinary differential equations. Let I be a non-degenerate interval in $\mathbf{R}$; given any $n \in \mathbf{N}$, let $f^{(n)}$ be the nth derivative of f and put

$$C^n(I) = \left\{f \in C(I) : f^{(n)} \text{ exists and is continuous on } I\right\}.$$

By a **linear ordinary differential equation of order n on I** we shall mean a problem, denoted by

$$x^{(n)} + a_1(t)x^{(n-1)} + \ldots + a_n(t)x = h(t), \tag{2.3.3}$$

of the following type:

Given $a_1, \ldots, a_n, h \in C(I)$, does there exist a function $x \in C^n(I)$ such that for all $t \in I$,

$$x^{(n)}(t) + a_1(t)x^{(n-1)}(t) + \ldots + a_n(t)x(t) = h(t) \text{ ?}$$

If there is such a function, it is called a **solution** of (2.3.3). Equation (2.3.3) is called **homogeneous** or **non-homogeneous** according to whether $h = 0$ or $h \neq 0$. Given $t_0 \in I$, an **initial-value problem set at t_0, associated with** (2.3.3), is the problem of whether given $(\eta_1, \ldots, \eta_n) \in \mathbf{R}^n$, there is a solution ϕ of (2.3.3) such that

$$\left(\phi(t_0), \phi^{(1)}(t_0), \ldots, \phi^{(n-1)}(t_0)\right) = (\eta_1, \ldots, \eta_n).$$

This problem is symbolised by

$$\left.\begin{aligned} x^{(n)} + a_1(t)x^{(n-1)} + \ldots + a_n(t)x \qquad\qquad &= h(t), \\ x(t_0) = \eta_1,\ x^{(1)}(t_0) = \eta_2,\ \ \ldots,\ x^{(n-1)}(t_0) &= \eta_n. \end{aligned}\right\} \tag{2.3.4}$$

Example 2.3.33

(i) Let $I = \mathbf{R}$. The problem

$$\ddot{x} - x = 0 \tag{2.3.5}$$

is a homogeneous ordinary differential equation of order 2. The function $t \longmapsto e^t$ is a solution. The problem

$$\ddot{x} - x = 0,\ x(0) = 1,\ \dot{x}(0) = 0$$

is an initial-value problem set at 0, associated with (2.3.5). Its unique solution is $t \longmapsto \cosh t$.

(ii) Let $I = (0, \infty)$. The problem

$$\ddot{x} + t^{-1}\dot{x} - t^{-2}x = \log\ t, \tag{2.3.6}$$

is an inhomogeneous equation of order 2, with $t \longmapsto (3\log t - 4)t^2/9$ as a solution. The problem

$$\ddot{x} + t^{-1}\dot{x} - t^{-2}x = \log\ t,\ x(1) = 1,\ \dot{x}(1) = 0$$

is an initial-value problem set at 1 associated with (2.3.6), and has the unique solution

$$t \longmapsto t + \frac{4}{9}t^{-1} + \frac{1}{9}(3\log t - 4)t^2.$$

We shall now prove the existence and uniqueness of solutions of initial-value problems, when $n = 2$ and I is closed and bounded.

Consider the initial-value problem

$$\left.\begin{aligned} \ddot{x} + a_1(t)\dot{x} + a_2(t)x &= h(t), \\ x(t_0) = \eta_0,\ \dot{x}(t_0) \quad &= \eta_1, \end{aligned}\right\} \tag{2.3.7}$$

where $a_1, a_2, h \in C(I)$ and $t_0 \in I$. Suppose $\phi \in C^2(I)$ is a solution of (2.3.7) and let $u = \ddot{\phi}$. By Taylor's theorem with the integral form of the remainder,

$$\phi(t) = \eta_0 + (t - t_0)\eta_1 + \int_{t_0}^{t} (t - s)u(s)ds,\ \dot{\phi}(t) = \eta_1 + \int_{t_0}^{t} u(s)ds.$$

Substitution in (2.3.7) now shows that

$$u(t) = h(t) - \eta_1 a_1(t) - \{\eta_0 + (t - t_0)\eta_1\}\, a_2(t) - \int_{t_0}^{t} \{a_1(t) + (t - s)a_2(t)\}\, u(s)ds,$$

and so u satisfies the **integral equation** (of **Volterra** type)

$$u(t) = g(t) + \int_{t_0}^{t} k(t, s)u(s)ds, \tag{2.3.8}$$

where

$$g(t) = h(t) - \eta_1 a_1(t) - \{\eta_0 + (t - t_0)\eta_1\}\, a_2(t),$$

$$k(t, s) = -\{a_1(t) + (t - s)a_2(t)\}\,.$$

Thus the second derivative of any solution of (2.3.7) is a solution of (2.3.8).

On the other hand, if the integral equation (2.3.8) has a unique solution $w \in C(I)$, let

$$\psi(t) = \eta_0 + (t - t_0)\eta_1 + \int_{t_0}^{t} (t - s)w(s)ds \quad (t \in I)$$

so that

$$\dot{\psi}(t) = \eta_1 + \int_{t_0}^{t} w(s)ds,\ \ddot{\psi}(t) = w(t)(t \in I).$$

Plainly ψ satisfies (2.3.7). Moreover, it is the unique such solution, for if ψ_1 were another, then $\ddot{\psi}_1 = w = \ddot{\psi}$ and so, by Taylor's theorem, $\psi_1 = \psi$.

It follows that the problem of the existence of a unique solution of (2.3.7) can be reduced to that of the existence of a unique solution of the integral equation (2.3.8). We now prove, with the aid of the contraction mapping theorem, that (2.3.8) does indeed have a unique solution.

Theorem 2.3.34 *Let I be closed and bounded, let $g \in C(I)$, $t_0 \in I$, put $D = I \times I$ and let $k : D \to \mathbf{R}$ be continuous. Then there is a unique $\phi \in C(I)$ satisfying the Volterra equation*

$$\phi(t) = g(t) + \int_{t_0}^{t} k(t,s)\phi(s)ds \quad (t \in I).$$

Proof Let $I = [a,b]$, $u \in C(I)$ and define ψ by

$$\psi(t) = g(t) + \int_{t_0}^{t} k(t,s)u(s)ds \quad (t \in I).$$

We claim that $\psi \in C(I)$. To prove this, first note that for fixed $t \in I$, the map $s \longmapsto k(t,s)u(s)$ belongs to $C(I) \subset \mathscr{R}(I)$. Now let $t_1 \in I$ and $\varepsilon > 0$. For each $t \in I$,

$$\begin{aligned} |\psi(t) - \psi(t_1)| &\le |g(t) - g(t_1)| + \left| \int_{t_0}^{t} k(t,s)u(s)ds - \int_{t_0}^{t_1} k(t_1,s)u(s)ds \right| \\ &\le |g(t) - g(t_1)| + \left| \int_{t_1}^{t} k(t,s)u(s)ds \right| \\ &\quad + \left| \int_{t_0}^{t_1} \{k(t,s) - k(t_1,s)\}\, u(s)ds \right|. \end{aligned}$$

Let

$$m = \sup\{|u(s)| : s \in I\}, \; M = \sup\{|k(t,s)| : (t,s) \in D\}.$$

In view of Theorem 2.3.16 and Corollary 2.3.25, both m and M are finite. The continuity of g at t_1 and the uniform continuity of k on the compact set D (see Theorem 2.3.30) imply that there exists $\delta > 0$ such that

$$|g(t) - g(t_1)| < \varepsilon/3, \; Mm\,|t - t_1| < \varepsilon/3 \text{ and } m\,|k(t,s) - k(t_1,s)|\,(b-a) < \varepsilon/3$$

if $s, t \in I = [a,b]$ and $|t - t_1| < \delta$. It follows that

$$|\psi(t) - \psi(t_1)| < \varepsilon \text{ if } t \in I \text{ and } |t - t_1| < \delta,$$

and so ψ is continuous on I.

Next, define $T : C(I) \to C(I)$ by

$$Tu(t) = g(t) + \int_{t_0}^{t} k(t,s)u(s)ds \; (t \in I, \; u \in C(I)).$$

Since $C(I)$ is a complete metric space when equipped with the uniform metric d_∞, we claim that for some $k \in \mathbf{N}$, T^k is a contraction mapping and so propose to use Corollary 2.2.14 to show that T has a unique fixed point. For each $n \in \mathbf{N}_0$, let $P(n)$ be the proposition

$$\left|T^n u(t) - T^n v(t)\right| \leq (M\,|t - t_0|)^n \, d_\infty(u,v)/n! \text{ for all } u, v \in C(I) \text{ and all } t \in I.$$

Evidently $P(0)$ is true; and if $P(n)$ is true for some $n \in \mathbf{N}_0$, then

$$\begin{aligned}\left|T^{n+1}u(t) - T^{n+1}v(t)\right| &= \left|\int_{t_0}^{t} k(t,s)\left\{T^n u(s) - T^n v(s)\right\} ds\right| \\ &\leq M \left|\int_{t_0}^{t} (M\,|s - t_0|)^n \,/n! ds\right| d_\infty(u,v) \\ &\leq (M\,|t - t_0|)^{n+1}\, d_\infty(u,v)/(n+1)!\end{aligned}$$

for all $u, v \in C(I)$ and all $t \in I$, so that $P(n+1)$ is true. Hence $P(n)$ is true for all $n \in \mathbf{N}$. Thus

$$d_\infty(T^n u, T^n v) \leq \frac{(M(b-a))^n}{n!} d_\infty(u,v)$$

for all $u, v \in C(I)$ and all $n \in \mathbf{N}$. Choose $k \in \mathbf{N}$ so large that $(M(b-a))^k\,/k! < 1$; T^k is a contraction. Hence by Corollary 2.2.14, T has a unique fixed point, ϕ say, and

$$\phi(t) = T\phi(t) = g(t) + \int_{t_0}^{t} k(t,s)\phi(s)ds (t \in I).$$

The proof is complete. □

As an immediate consequence of this theorem we have

Corollary 2.3.35 *Let I be closed and bounded. Then the initial-value problem (2.3.7) has a unique solution.*

Next we show how the Arzelà-Ascoli theorem may be used to prove a famous theorem, due to Peano , which establishes the existence of a solution of the initial-value problem for a non-linear differential equation.

Theorem 2.3.36 *Let $t_0, x_0 \in \mathbf{R}$ and $a, b > 0$, put $I = [t_0, t_0+a]$, $J = [x_0-b, x_0+b]$ and suppose that $f : I \times J \to \mathbf{R}$ is continuous, with*

$$M = \max_{(t,x)\in I\times J} |f(t,x)| > 0;$$

put $c = \min(a, b/M)$. Then there is a function $x \in C^1([t_0, t_0+c])$ such that

$$\dot{x}(t) = f(t, x(t)) \text{ for } t \in [t_0, t_0 + c], x(t_0) = x_0. \tag{2.3.9}$$

Proof Plainly x is a solution of (2.3.9) if, and only if,

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds, \;\; t \in [t_0, t_0 + c]. \tag{2.3.10}$$

For simplicity of exposition, suppose that $t_0 = 0$; the general case is handled similarly. Put $I_1 = [0, c]$ and for each $n \in \mathbf{N}$ define $x_n : I_1 \to \mathbf{R}$ by

$$x_n(t) = \begin{cases} x_0, & 0 \leq t \leq c/n, \\ x_0 + \int_0^{t-c/n} f(s, x_n(s))ds, & c/n < t \leq c. \end{cases}$$

The function x_n is well-defined: it is given by

$$x_n(t) = x_{j,n}(t) \text{ for } jc/n \leq t \leq (j+1)c/n \text{ and } j = 0, 1, \ldots, n-1,$$

where

$$x_{0,n}(t) = x_0 \;\; (0 \leq t \leq c/n),$$
$$x_{1,n}(t) = x_0 + \int_0^{t-c/n} f(s, x_0)ds \;\; (c/n < t \leq 2c/n)$$

and, for $j = 2, \ldots, n-1$ and $jc/n < t \leq (j+1)c/n$,

$$x_{j,n}(t) = x_0 + \sum_{k=1}^{j-1} \int_{(k-1)c/n}^{kc/n} f(s, x_{k-1,n}(s))ds + \int_{(j-1)c/n}^{t-c/n} f(s, x_{j-1,n}(s))ds.$$

It is clear that $x_n \in C(I_1)$. Moreover, for all $t \in I_1$ and all $n \in \mathbf{N}$,

$$|x_n(t) - x_0| \leq cM \leq b \text{ and } |x_n(t)| \leq |x_0| + b.$$

Hence the sequence (x_n) is uniformly bounded. In fact, it is equicontinuous, for given any $n \in \mathbf{N}$ and any $t_1, t_2 \in I_1$,

$$|x_n(t_1) - x_n(t_2)| \leq \left| \int_{t_1 - c/n}^{t_2 - c/n} f(s, x_n(s))ds \right| \leq M \left| t_2 - t_1 \right|.$$

Hence by the Arzelà-Ascoli theorem (Theorem 2.3.22), there is a subsequence $(x_{k(n)})$ of (x_n) which is uniformly convergent on I_1, to x say. For all $t \in I_1$, as $k(n) \to \infty$,

$$\left|x_{k(n)}(t) - x_0 - \int_0^t f(s, x_{k(n)}(s))ds\right| = \left|\int_{t-c/k(n)}^t f(s, x_{k(n)}(s))ds\right|$$
$$\leq Mc/k(n) \to 0.$$

Since f is uniformly continuous on the compact set $I_1 \times J$, $f(s, x_{k(n)}(s))$ converges uniformly on $I_1 \times J$ to $f(s, x(s))$, and

$$\int_0^t f(s, x_{k(n)}(s))ds \to \int_0^t f(s, x(s))ds$$

as $k(n) \to \infty$. Thus

$$x(t) = x_0 + \int_0^t f(s, x(s))ds, \ t \in I_1,$$

and the proof is complete. □

Note that there may well be more than one solution of the initial-value problem (2.3.9). For example, the initial-value problem

$$\dot{x}(t) = |x(t)|^{1/2} \text{ for } t \in [0, 1], \ x(0) = 0,$$

has, apart from the zero function, a whole family of solutions given by

$$x(t) = \begin{cases} 0, & 0 \leq t \leq c, \\ (t-c)^2/4, & c < t \leq 1, \end{cases}$$

for any $c \in (0, 1)$. Sufficient conditions on the function f for uniqueness to be restored are given in Exercise 2.3.38/14 below.

2.3.2 *Application 2*

Here we revisit the Riemann integral and give a celebrated criterion for functions to be Riemann-integrable. To do this, we need the concept of a null set. A subset E of $\mathbf{R}$ is said to be a **null set** if, given any $\varepsilon > 0$, there is a sequence (I_n) of intervals I_n of length $l(I_n)$ such that $E \subset \cup_n I_n$ and $\sum_{n=1}^{\infty} l(I_n) < \varepsilon$. It is clear that every finite set is a null set, as is every subset of a null set. Somewhat less obviously, if (E_n) is a sequence of null sets, then $\cup_{n=1}^{\infty} E_n$ is a null set. To establish this, let $\varepsilon > 0$ and note that given any $n \in \mathbf{N}$, there is a sequence $(I_m^{(n)})$ of intervals such that $E_n \subset \cup_{m=1}^{\infty} I_m^{(n)}$, $\sum_{m=1}^{\infty} l(I_m^{(n)}) < \varepsilon/2^n$. The sequence $(I_m^{(n)})_{m,n\in\mathbf{N}}$ is countable and so may be arranged as a sequence $(J_k)_{k\in\mathbf{N}}$, with $\cup_{n=1}^{\infty} E_n \subset \cup_{k=1}^{\infty} J_k$, and

$$\sum_{k=1}^{\infty} l(J_k) \leq \sum_{n=1}^{\infty} \varepsilon/2^n = \varepsilon.$$

This justifies our claim.

The criterion mentioned above is as follows.

Theorem 2.3.37 *Let $a, b \in \mathbf{R}$, with $a < b$, let $f \in \mathscr{B}[a,b]$ and set*

$$D_f = \{x \in [a,b] : f \text{ is not continuous at } x\}.$$

Then $f \in \mathscr{R}[a,b]$ if, and only if, D_f is a null set.

Proof We may clearly suppose that f is not the zero function. Let $M = \sup\{|f(x)| : x \in [a,b]\}$ and for each $n \in \mathbf{N}$ put

$$E_n = \{x \in [a,b] : \text{ for all } \delta > 0 \text{ there exist } s, t \in (x-\delta, x+\delta) \cap [a,b] \text{ such that } |f(s) - f(t)| > 1/n\}.$$

Plainly f is not continuous at x if $x \in E_n$ for some n. On the other hand, if $x \in D_f$, then there is a sequence (x_k) in $[a,b]$, with $x_k \to x$, such that $f(x_k) \nrightarrow f(x)$. This implies that there exist $n \in \mathbf{N}$ and a subsequence of (x_k), still denoted by (x_k) for convenience, such that $|f(x_k) - f(x)| > 1/n$ for all $k \in \mathbf{N}$. Thus $x \in E_n$. It follows that

$$D_f = \bigcup_{n \in \mathbf{N}} E_n.$$

We claim that each E_n is compact. Since E_n is obviously bounded, it is sufficient to prove that it is closed. To do this, let $x \in \overline{E_n}$. Given $\delta > 0$, there exists $y \in E_n$ with $|x - y| < \delta/2$; and since $(y - \delta/2, y + \delta/2) \subset (x - \delta, x + \delta)$ and there are $s, t \in (y - \delta/2, y + \delta/2)$ with $|f(s) - f(t)| > 1/n$, it follows that $x \in E_n$, which establishes our claim.

Now suppose that $f \in \mathscr{R}[a,b]$. By Exercise 1.1.10/7, given $n \in \mathbf{N}$ and $\varepsilon > 0$, there is a partition $P = \{a = x_0, x_1, \ldots, x_m = b\} \in \mathscr{P}[a,b]$ such that

$$\left|\sum_{r=1}^{m} \{f(\xi_r) - f(\eta_r)\}(x_r - x_{r-1})\right| < \varepsilon/n$$

whenever $\xi_r, \eta_r \in [x_{r-1}, x_r]$, for $r \in \{1, \ldots, m\}$. For each $r \in \{1, \ldots, m\}$ we may plainly choose ξ_r, η_r so that $f(\xi_r) \geq f(\eta_r)$; moreover, if $(x_{r-1}, x_r) \cap E_n \neq \emptyset$, we may ensure that $f(\xi_r) > f(\eta_r) + 1/n$. It now follows that the sum of the lengths of those intervals (x_{r-1}, x_r) with non-empty intersection with E_n is less than ε. Hence, since the length of degenerate intervals is zero, E_n is a null set; and as D_f is the countable union of the E_n, it also is a null set.

For the converse, suppose that D_f is a null set. Let $\varepsilon > 0$ and choose $n \in \mathbf{N}$ so that $n > 1/\varepsilon$. Since E_n is obviously null, there is a sequence (I_r) of open subintervals of the metric space $[a,b]$ which covers E_n, with $\sum_{r=1}^{\infty} l(I_r) < \varepsilon$. As E_n is compact,

it is covered by a finite number of these intervals, say $J_1, \ldots, J_p$; and of course $\sum_{r=1}^{p} l(J_r) < \varepsilon$. An inductive argument shows that the set

$$[a, b] \setminus \bigcup_{r=1}^{p} J_r$$

consists of a finite collection of closed intervals, say $K_1, \ldots, K_q$; for each $j \in \{1, \ldots, q\}$, there exists $P_j \in \mathscr{P}(K_j)$ such that $|f(x) - f(y)| \leq 1/n$ for all x, y in the same subinterval of P_j. Finally, let $P \in \mathscr{P}[a, b]$ consist of the points of $\bigcup_{j=1}^{q} P_j$ together with the endpoints of the intervals $J_1, \ldots, J_p$. Then, using Exercise 1.1.10/2, we see that the contribution to $U(P,f) - L(P,f)$ from the points of $P_1, \ldots, P_q$ can be estimated from above by

$$\frac{1}{n}(b - a) < \varepsilon(b - a).$$

The rest of $U(P,f) - L(P,f)$ arises from the endpoints of the J_r and may be estimated from above by

$$2M \sum_{r=1}^{p} l(J_r) < 2M\varepsilon.$$

Hence

$$U(P,f) - L(P,f) < \varepsilon(2M + b - a),$$

and so $f \in \mathscr{R}[a, b]$. □

Note that this theorem gives an immediate proof of the fact, established earlier, that Riemann-integrability is preserved by taking sums and products.

Exercise 2.3.38

1. Let $I = [0, 1]$. Exhibit a subset of the metric space $C(I)$, endowed with the uniform metric, that is unbounded. Show that the mapping $f \longmapsto \int_0^1 f$ of $C(I)$ to $\mathbf{R}$ is uniformly continuous on $C(I)$.
2. Let (X, d) be a compact metric space and let $(F_i)_{i \in I}$ be a family of non-empty closed subsets of X with empty intersection. Prove that there is a positive number c such that for each $x \in X$, $d(x, F_i) \geq c$ for some $i \in I$.
3. Let (X, d) be a compact metric space such that for all $x, y, z \in X$, $d(x, y) \leq \max\{d(x, z), d(y, z)\}$, and let $x_0 \in X$; let $x \in X$ be such that $d(x_0, x) = r > 0$. By assuming the contrary show that

   $$\sup\{d(x_0, y) : y \in B(x_0, r)\} < r \text{ and } \inf\{d(x_0, y) : y \in X, d(x_0, y) > r\} > r.$$

   Hence prove that $\{d(x_0, z) : z \in X\}$ is finite or countably infinite.
4. Let (X, d) be a compact metric space, let $T : X \to X$ be such that for all $x, y \in X$, $d(x, y) \leq d(T(x), T(y))$, and let a, b be any points of X. By considering appropriate subsequences of $(T^n(a))$ and $(T^n(b))$, show that given any $\varepsilon > 0$, there is an integer k such that $d(a, T^k(a)) < \varepsilon$ and $d(b, T^k(b)) < \varepsilon$. Deduce that

$d(T(a), T(b)) = d(a, b)$ and that $T(X)$ is dense in X. Hence show that T maps X isometrically onto itself.

5. Let (X, d) be a compact metric space and suppose that $T : X \to X$ is such that $d(T(x), T(y)) < d(x, y)$ for all $x, y \in X$ with $x \neq y$. Prove that T has a unique fixed point.
6. (Dini's theorem: see also Exercise 1.7.17/18) Let (X, d) be a compact metric space and let (f_n) be a monotone sequence in $C(X)$ which is pointwise convergent to $f \in C(X)$. Prove that $f_n \to f$ in the uniform metric on $C(X)$.
7. Let $\alpha \in (0, 1]$. A real-valued function f on $[0, 1]$ is said to be Hölder-continuous with exponent α if there is a constant C such that for all $x, y \in [0, 1]$, $|f(x) - f(y)| \leq C\,|x - y|^{\alpha}$. Define

$$\|f\|_{\alpha} = \max_{x\in[0,1]} |f(x)| + \sup \frac{|f(x) - f(y)|}{|x - y|^{\alpha}},$$

where the supremum is taken over all $x, y \in [0, 1]$ with $x \neq y$. Prove that the set of all functions f with $\|f\|_{\alpha} \leq 1$ is a compact subset of $C[0, 1]$.
8. Let $\mathscr{K} = \{f \in C[0, 1] : d_{\infty}(f, 0) \leq 1\}$. Show that $\mathscr{K}$ is not compact in $C[0, 1]$.
9. Let (X, d) be a compact metric space and let (f_n) be a sequence in $C(X)$. Prove that if the set $\{f_n : n \in \mathbf{N}\}$ is equicontinuous, and for each $x \in X$ the sequence $(f_n(x))$ converges, then (f_n) is convergent in $C(X)$.
10. Let $f_n(t) = \sin\sqrt{t + 4n^2\pi^2}$ for $t \geq 0$, $n \in \mathbf{N}$. Prove that $\{f_n : n \in \mathbf{N}\}$ is a bounded and uniformly equicontinuous subset of $\mathscr{C}[0, \infty)$, but that it is not relatively compact. Prove also that the sequence (f_n) converges pointwise to 0 on $[0, \infty)$. [This shows that the Arzelà-Ascoli theorem and Exercise 9 may fail when X is not compact.]
11. Let $\mathscr{K} \subset C[0, 1]$. Suppose that each $f \in \mathscr{K}$ is differentiable on $(0, 1)$ and that there exists $M > 0$ such that $\left|f'(t)\right| \leq M$ for all $t \in (0, 1)$ and all $f \in \mathscr{K}$. Prove that $\mathscr{K}$ is equicontinuous.
12. Let X be a metric space, $x \in X$, and f be a real-valued function on X. Prove that f is lower semi-continuous at $x \in X$ if, and only if,

$$f(x) \leq \liminf_{n\to\infty} f(x_n) \text{ whenever } x_n \to x.$$

13. Let (X, d) be a metric space and let $f : X \to \mathbf{R}$ be bounded and lower semi-continuous. For each $n \in \mathbf{N}$, let $g_n : X \to \mathbf{R}$ be defined by

$$g_n(x) = \inf_{y\in X}\{f(y) + nd(x, y)\} \ (x \in X).$$

(i) Prove that (g_n) is an increasing sequence of continuous functions that converges pointwise to f.
(ii) Show that the set of points of continuity of f is residual in X and deduce that, if X is complete, then this set is dense in X.

14. Let $a, b, c \in \mathbf{R}$ with $a < b$ and $c \geq 0$, let u, v be non-negative continuous functions on $[a, b]$ and suppose that

$$v(t) \leq c + \int_a^t v(s)u(s)ds \text{ for } a \leq t \leq b.$$

Establish Gronwall's inequality:

$$v(t) \leq c \exp\left(\int_a^t u(s)ds\right) \text{ for } a \leq t \leq b,$$

so that if $c = 0$, then v is the zero function. Deduce that the initial-value problem (2.3.9) has a unique solution if the function f is Lipschitz-continuous in the sense that there is a constant K such that

$$|f(t, w_1) - f(t, w_2)| \leq K\,|w_1 - w_2| \text{ for all } t \in [t_0, t_0 + c] \text{ and all } w_1, w_2 \in J.$$

15. Let $\mathscr{U}$ be an open covering of a compact metric space X. Show that there is a positive number ε (called a Lebesgue number of $\mathscr{U}$) such that if $A \subset X$ and $\operatorname{diam} A < \varepsilon$, then there exists $U \in \mathscr{U}$ that contains A.
16. Let (X, d) be a complete metric space and let $\mathscr{K}$ be the family of all non-empty compact subsets of X. The Hausdorff metric δ on $\mathscr{K}$ is defined by

$$\delta(A, B) = \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\}\ (A, B \in \mathscr{K}),$$

in the notation of Lemma 2.1.40. Show that

$$\delta(A, B) = \inf\{r > 0 : A \subset V_r(B), B \subset V_r(A)\},$$

where $V_r(A) = \{x \in X : d(x, A) < r\}$. Prove that δ is a metric on $\mathscr{K}$ and that $(\mathscr{K}, \delta)$ is complete. Show further that if X is compact, then so is $(\mathscr{K}, \delta)$. Prove that if for each $i \in \{1, \ldots, n\}$, A_i and B_i belong to $\mathscr{K}$, then

$$\delta\left(\cup_{i=1}^n A_i, \cup_{i=1}^n B_i\right) \leq \max_{1 \leq i \leq n} \delta(A_i, B_i).$$

Let $F : X \to X$ be a contraction; that is, there exists $r \in (0, 1)$ such that for all $x, y \in X$, $d(F(x), F(y)) \leq rd(x, y)$. Prove that for all $A, B \in \mathscr{K}$,

$$\delta(F(A), F(B)) \leq r\delta(A, B).$$

Now suppose that for each $i \in \{1, \ldots, n\}$, $F_i : X \to X$ is a contraction. Define $\mathscr{F} : \mathscr{K} \to \mathscr{K}$ by $\mathscr{F}(A) = \cup_{i=1}^n F_i(A)$ $(A \in \mathscr{K})$, show that $\mathscr{F}$ is a contraction on $(\mathscr{K}, \delta)$ and hence prove that there is a unique $K \in \mathscr{K}$ such that

$$K = \cup_{i=1}^{n} F_i(K).$$

By taking $X = [0, 1]$ (with the metric inherited from $\mathbf{R}$), $n = 2$, $F_1(x) = x/3$ and $F_2(x) = (2 + x)/3$ $(x \in [0, 1])$, deduce that $\lim_{n\to\infty} \mathscr{F}^n([0, 1])$ exists in $(\mathscr{K}, \delta)$ and so defines a compact non-empty subset of $[0, 1]$. This is the Cantor set.

2.4 Connectedness

In this section we isolate those metric spaces with the following property: if a map $f : X \to \mathbf{R}$ is continuous, then its range, $f(X)$, is an interval. The motivation for this stems from the well-known intermediate-value theorem.

We begin with a characterisation of those subsets of $\mathbf{R}$ which are intervals.

Lemma 2.4.1 *A subset S of $\mathbf{R}$ is an interval if, and only if, it has the following **intermediate-value property** (abbreviated as ivp):*

$$\textit{if } x, y \in S \textit{ and } x < z < y, \textit{ then } z \in S.$$

Proof If S has at most one element, it is a degenerate interval and the result holds by default.

Suppose that S has at least two elements. If it is an interval then it clearly has the ivp. To establish the converse we distinguish four cases:

(i) $\inf S = a > -\infty$, $\sup S = b < \infty$. Evidently $S \subset [a, b]$; we claim that $(a, b) \subset S$. For suppose that $x \in (a, b)$. Then there exist $c, d \in S$ such that $a \le c < x < d \le b$ and hence, by the ivp, $x \in S$. Thus $(a, b) \subset S \subset [a, b]$ and S is an interval.
(ii) $\inf S = a > -\infty$, $\sup S = \infty$. Here $S \subset [a, \infty)$. If $x \in (a, \infty)$, then there are $c, d \in S$ such that $a \le c < x < d < \infty$ and, as before, $x \in S$. Thus $(a, \infty) \subset S \subset [a, \infty)$ and S is an interval.
(iii) $\inf S = -\infty$, $\sup S = b < \infty$.
(iv) $\inf S = -\infty$, $\sup S = \infty$.

We omit the proofs in cases (iii) and (iv) as they are similar to that of case (ii). □

We can now give equivalent forms of the property with which we began this section.

Theorem 2.4.2 *Let X be a metric space. The following three statements are equivalent:*

(i) *The only subsets of X which are both open and closed are $\emptyset$ and X.*
(ii) *There do not exist two non-empty disjoint open subsets of X whose union is X.*
(iii) *The range of each continuous map $f : X \to \mathbf{R}$ is an interval.*

Proof Suppose that (i) holds and that (ii) does not. Then there are non-empty open subsets U, V of X such that $U \cap V = \emptyset$ and $U \cup V = X$. This implies that $U = {}^cV$ and so U is closed. Thus $\emptyset \neq U \neq X$ and U is both open and closed, contradicting (i).

Now suppose that (ii) holds and (iii) does not. Then there is a continuous map $f : X \to \mathbf{R}$ such that $f(X)$ is not an interval. Hence, in view of Lemma 2.4.1, there exist $x, y \in X$ and $\lambda \in \mathbf{R}$ such that $f(x) < \lambda < f(y)$ and, for all $z \in X, f(z) \neq \lambda$. Let $U = f^{-1}((-\infty, \lambda))$ and $V = f^{-1}((\lambda, \infty))$. These sets are non-empty, disjoint, open and their union is X, contradicting (ii).

Finally, suppose (iii) holds and (i) does not. Then there is a set U which is both open and closed in X and $\emptyset \neq U \neq X$. Define $f : X \to \mathbf{R}$ by $f(x) = 1$ if $x \in U$, $f(x) = 0$ otherwise. Since $f^{-1}(W) \in \{\emptyset, U, {}^cU, X\}$ if $W \subset \mathbf{R}$, it follows that $f^{-1}(W)$ is open in X whenever W is open in $\mathbf{R}$. Hence f is continuous, but its range is not an interval and (iii) is contradicted. $\square$

This leads us to formulate the following definition.

Definition 2.4.3 A metric space X is said to be **connected** if it is not expressible as a union of two non-empty, disjoint open subsets of itself; it is said to be **disconnected** if it is not connected.

Of course, any of the equivalences of Theorem 2.4.2 could have been used to define a connected space. There is some loss of motivation in not choosing (iii), but the compensation is that we have an intrinsic and functional definition.

We now turn to subsets of a metric space.

Definition 2.4.4 A subset of a metric space X is said to be a **connected set** in X if it is either empty or it is connected as a subspace of X; it is said to be a **disconnected set** in X if it is not a connected set in X.

Let E be a subspace of a metric space X. By definition, E is a disconnected space if, and only if, there are non-empty sets O_1 and O_2, each open in E, such that $O_1 \cap O_2 = \emptyset$ and $O_1 \cup O_2 = E$. If $\mathscr{U}$ denotes the family of all the sets open in X, then $\{U \cap E : U \in \mathscr{U}\}$ is the family of all the sets open in the metric space E. It follows that E is a disconnected space if, and only if, there are sets U and V, each open in X, such that

$$U \cap E \neq \emptyset, \ V \cap E \neq \emptyset$$

and

$$(U \cap E) \cap (V \cap E) = \emptyset, \ (U \cap E) \cup (V \cap E) = E.$$

With the observation that
(a) $(U \cap E) \cap (V \cap E) = \emptyset$ if, and only if, $U \cap V \cap E = \emptyset$,
and
(b) $(U \cap E) \cup (V \cap E) = E$ if, and only if, $E \subset U \cup V$,
this means that we have established the following theorem.

Theorem 2.4.5 *Let E be a subset of a metric space X. Then E is a disconnected set in X if, and only if, there are sets U and V, each open in X, such that*

$$U \cap E \neq \emptyset, \ V \cap E \neq \emptyset,$$

$$U \cap V \cap E = \emptyset, \ E \subset U \cup V.$$

In practice, given a set E in a metric space X, Theorem 2.4.5 provides a basic test for its disconnectedness. In the event that the set E is known to be disconnected, a condition stronger in form than the test-condition of Theorem 2.4.5 holds. This appears next.

Theorem 2.4.6 *Let E be a subset of a metric space X. Then E is a disconnected set in X if, and only if, there are* ***disjoint*** *open sets U and V in X such that $U \cap E \neq \emptyset$, $V \cap E \neq \emptyset$ and $E \subset U \cup V$.*

Proof Let E be disconnected in X. Then there are sets U_1 and V_1, each open in X, such that $E \cap U_1 \neq \emptyset$, $E \cap V_1 \neq \emptyset$, $E \cap U_1 \cap V_1 = \emptyset$ and $E \subset U_1 \cup V_1$. Moreover, given any $u \in E \cap U_1$, there exists $r(u) > 0$ such that $B(u, r(u)) \subset U_1$; also, given any $v \in E \cap V_1$, there exists $r(v) > 0$ such that $B(v, r(v)) \subset V_1$. Put

$$U = \bigcup_{u \in E \cap U_1} B(u, r(u)/2), \ V = \bigcup_{v \in E \cap V_1} B(v, r(v)/2).$$

It is clear that U and V are open, that $E \cap U = E \cap U_1 \neq \emptyset$ and $E \cap V = E \cap V_1 \neq \emptyset$, and that $E = (E \cap U_1) \cup (E \cap V_1) \subset U \cup V$. It remains to prove that $U \cap V = \emptyset$. To obtain a contradiction, suppose that $U \cap V \neq \emptyset$. Let $w \in U \cap V$. Then there are points $u \in E \cap U_1$, $v \in E \cap V_1$ such that $d(u, w) < \frac{1}{2} r(u)$, $d(v, w) < \frac{1}{2} r(v)$, where d is the metric on X. Thus

$$d(u, v) \leq d(u, w) + d(w, v) \leq \frac{1}{2} \{r(u) + r(v)\} \leq \max \ \{r(u), r(v)\}.$$

It follows that either $v \in U_1$ or $u \in V_1$. Whichever is the case, $U_1 \cap V_1 \cap E \neq \emptyset$, and we have a contradiction.

The converse is obvious. □

Example 2.4.7

(i) In every metric space (X, d) any set containing only one point is obviously connected; any finite set with at least two points is disconnected. Thus if $S = \{a, b\} \subset X$ and $a \neq b$, for example, we may take $U = B(a, r)$, $V = B(b, r)$, where $r = \frac{1}{2} d(a, b)$, and note that U and V are open, $U \cap V = \emptyset$, $U \cap S \neq \emptyset$, $V \cap S \neq \emptyset$ and $S \subset U \cup V$.

(ii) In any discrete metric space every subset with more than one point is disconnected, as every subset of the space is both open and closed.

(iii) Let X be a metric space, let A, B be non-empty, disjoint, closed sets in X and let $E = A \cup B$. Then E is disconnected. To see this, put $U = {}^cA$, $V = {}^cB$ so that U and V are open in X. Then $U \cap E = B \neq \emptyset$, $V \cap E = A \neq \emptyset$, $U \cap V \cap E = {}^cA \cap {}^cB \cap (A \cup B) = (A \cup B)^c \cap (A \cup B) = \emptyset$ and $E = A \cup B \subset {}^cB \cup {}^cA = U \cup V$. To illustrate this, let $X = \mathbf{R}^2$, $A = \{(x, y) \in \mathbf{R}^2 : x \geq 0, xy = 1\}$, $B = \{(x, y) \in \mathbf{R}^2 : y = 0\}$. Then $A \cup B$ is disconnected in $\mathbf{R}^2$.

(iv) Let X be a metric space and let A, B be non-empty, open, disjoint sets in X with union X. Then if C is a connected subset of X, either $C \subset A$ or $C \subset B$. For otherwise $C \cap A \neq \emptyset$, $C \cap B \neq \emptyset$, $C \cap A \cap B = \emptyset$ and $C \subset A \cup B$, and the connectedness of C is contradicted.

(v) A metric space X is connected if, and only if, given any $x, y \in X$, there is a connected subset A of X such that $x, y \in A$. To prove this, suppose first that given any $x, y \in X$, there is a connected subset A of X such that $x, y \in A$. If X were not connected, there would be disjoint, open, non-empty sets U, V with union X. By (iv), either $A \subset U$ or $A \subset V$, and we have a contradiction. The converse is obvious.

The connected subsets of $\mathbf{R}$, equipped with the usual metric, can be classified completely.

Theorem 2.4.8 *Let $S \subset \mathbf{R}$. The following three statements are equivalent.*

(i) *S is connected.*
(ii) *S has the intermediate-value property.*
(iii) *S is an interval.*

Proof Suppose that S is connected yet fails to have the intermediate-value property. Then there are real numbers x, y and z with $x, y \in S$, $x < z < y$ and $z \notin S$. Put $U = \{t \in \mathbf{R} : t < z\}$ and $V = \{t \in \mathbf{R} : t > z\}$. Then U and V are open, $S \cap U \neq \emptyset$, $S \cap V \neq \emptyset$, $U \cap V = \emptyset$ and $S \subset U \cup V$. Thus S is disconnected and we have a contradiction. Hence (i) implies (ii).

Conversely, suppose that S has the intermediate-value property and is disconnected. Then there are disjoint open sets U, V in $\mathbf{R}$ and points $x, y \in S$ with $x < y$ such that $x \in S \cap U$, $y \in S \cap V$ and $S \subset U \cup V$. Let $z := \sup\{U \cap [x, y]\}$. Plainly $z \in \overline{U}$ and, as $\overline{U}$ is contained in the closed set $\mathbf{R}\backslash V$, $z \notin V$. Since $z \in [x, y] \subset S \subset U \cup V$, it follows that $z \in U$. Since U is open and $z \neq y$, there exists $z_1 > z$ such that $[z, z_1] \subset U \cap [x, y]$. But this contradicts the definition of z. Hence (ii) implies (i).

The rest of the proof follows from Lemma 2.4.1. □

Corollary 2.4.9 *Let $S \subset \mathbf{R}$, $S \neq \emptyset$. Then S is an interval if, and only if, $f(S)$ has the ivp whenever $f : S \to \mathbf{R}$ is continuous. [The 'only if' part of this result is called the* ***intermediate-value*** *theorem.]*

Proof By Theorem 2.4.8, S is an interval if, and only if, S is connected; by Theorem 2.4.2, this is so if, and only if, $f(S)$ is an interval whenever $f : S \to \mathbf{R}$ is continuous; and now the result follows from Lemma 2.4.1. □

Corollary 2.4.10 *Let $a, b \in \mathbf{R}$, with $a < b$, and let $f : [a, b] \to [a, b]$ be continuous. Then f has a fixed point; that is, there exists $c \in [a, b]$ such that $f(c) = c$.*

Proof If $f(a) = a$ or $f(b) = b$ there is nothing to prove. We shall therefore assume that $f(a) > a$ and $f(b) < b$. Define $g : [a, b] \to \mathbf{R}$ by $g(x) = x - f(x)$, $x \in [a, b]$. Then g is continuous, $g(a) < 0$ and $g(b) > 0$. By Corollary 2.4.9, there exists $c \in [a, b]$ such that $g(c) = 0$; that is, $f(c) = c$. □

This elementary fixed-point result may be extended to higher dimensions with considerably greater effort: see Chap. 3 for the two-dimensional version.

Under a continuous map connectedness is preserved. Amongst other uses this fact allows new connected sets to be generated from old.

Theorem 2.4.11 *Let X and Y be metric spaces and let $f : X \to Y$ be continuous. Then $f(E)$ is a connected subset of Y whenever E is a connected subset of X.*

Proof Suppose that E is connected and yet $f(E)$ is not. Then there are disjoint open sets U, V in Y such that $U \cap f(E) \neq \emptyset$, $V \cap f(E) \neq \emptyset$ and $f(E) \subset U \cup V$. It follows that $f^{-1}(U) \cap E \neq \emptyset$, $f^{-1}(V) \cap E \neq \emptyset$, $E \subset f^{-1}(U) \cup f^{-1}(V)$ and, since $U \cap V = \emptyset$, $f^{-1}(U) \cap f^{-1}(V) = \emptyset$. As f is continuous, $f^{-1}(U)$ and $f^{-1}(V)$ are also open in X. Thus E is disconnected and we have a contradiction. □

Corollary 2.4.12 *Let $S = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = r^2\}$, where $r > 0$. Let $f : S \to \mathbf{R}$ be continuous (S inherits the Euclidean metric from $\mathbf{R}^2$). Then there exists $\mathbf{u} = (u, v) \in S$ such that $f(\mathbf{u}) = f(-\mathbf{u})$.*

Proof Note that S is connected: it is the image of the interval $[0, 2\pi]$ under the continuous map $t \longmapsto (r \cos t, r \sin t)$.

Let $g : S \to \mathbf{R}$ be defined by

$$g(\mathbf{p}) = f(\mathbf{p}) - f(-\mathbf{p}).$$

Then g is continuous: if $\mathbf{p}_n \in S \ \ (n \in \mathbf{N})$ and $\mathbf{p}_n \to \mathbf{p} \in S$, then $f(\mathbf{p}_n) \to f(\mathbf{p})$ and $f(-\mathbf{p}_n) \to f(-\mathbf{p})$, so that $g(\mathbf{p}_n) \to g(\mathbf{p})$. Since S is connected, it follows from Theorem 2.4.2 that $g(S)$ is an interval. This interval is symmetric about the origin: if $\theta \in g(S)$, then $g(\mathbf{s}) = \theta$ for some $\mathbf{s} \in S$ and so $-\theta = -g(\mathbf{s}) = g(-\mathbf{s}) \in g(S)$. Hence $0 \in g(S)$ and there exists $\mathbf{u} \in S$ with $g(\mathbf{u}) = 0$; that is, $f(\mathbf{u}) = f(-\mathbf{u})$. □

The use of the term connected in the context of metric spaces may seem remote from the everyday sense in which the term is employed. That sense, in which elements are linked or joined, does have a specialised counterpart for which the technical expression is path-connected. Three definitions introduce this.

Definition 2.4.13 Let X be a metric space and let $a, b \in \mathbf{R}$, with $a < b$. A continuous map $\gamma : [a, b] \to X$ is called a **path in** X with **parameter interval** $[a, b]$. The points $\gamma(a)$, $\gamma(b)$ are called the **initial** and **terminal** points, respectively, of γ; γ is said to **join** its initial and terminal points; γ is a **closed** path if $\gamma(a) = \gamma(b)$; γ is a **simple**

path if $\gamma(s) \neq \gamma(t)$ whenever $s, t \in [a, b]$, $s \neq t$ and $\{s, t\} \neq \{a, b\}$. The range $\gamma^* = \gamma([a, b])$ of γ is called the **track** of γ. If $\gamma^* \subset E \subset X$ we refer to γ as a **path in** E.

Without loss of generality, any path may be chosen to have parameter interval $[0, 1]$: make the obvious change of variable $t \mapsto (1-t)a + tb : [0, 1] \to [a, b]$.

Example 2.4.14 The function $\gamma : [0, 1] \to \mathbf{R}^2$ defined by $\gamma(t) = (\cos \pi t, \sin \pi t)$ is a path in $\mathbf{R}^2$ which joins its initial point $(1, 0)$ to its terminal point $(-1, 0)$ and has track

$$\gamma^* = \{(x, y) \in \mathbf{R}^2 : x^2 + y^2 = 1, y \geq 0\}.$$

Observe that different paths may have the same track: the path $\nu : [0, 1] \to \mathbf{R}^2$ given by $\nu(t) = (\cos \pi t^2, \sin \pi t^2)$ has the same track as γ, though $\nu \neq \gamma$.

Paths in $\mathbf{R}^n$ of a particular character are singled out.

Definition 2.4.15 Given $a, b \in \mathbf{R}$ with $a < b$, a map $\gamma : [a, b] \to \mathbf{R}^n$ is said to be a **polygonal path** if points $x^{(0)}, x^{(1)}, \ldots, x^{(k)} \in \mathbf{R}^n$ and a partition $\{a = t_0, t_1, \ldots, t_k = b\}$ of $[a, b]$ exist such that

$$\gamma(t) = (t_j - t_{j-1})^{-1} \left\{(t_j - t)x^{(j-1)} + (t - t_{j-1})x^{(j)}\right\}$$

whenever $t_{j-1} \leq t \leq t_j$ and $j \in \{1, 2, \ldots, k\}$; if, in addition, γ is such that for each $j \in \{1, 2, \ldots, k\}$ there is a line passing through $x^{(j-1)}$ and $x^{(j)}$ parallel to a coordinate axis, then it is said to be a **p-path**. In the elementary case of $k = 1$, when

$$\gamma(t) = (b - a)^{-1}\{(b - t)x^{(0)} + (t - a)x^{(1)}\} \ (a \leq t \leq b),$$

the path γ is referred to as a **line segment** and may be denoted by $[x^{(0)}, x^{(1)}]$. This terminology and symbolism is used also for γ^*, the track of γ, and the intended meaning has to be understood by context.

Elementary reasoning shows that a polygonal path is continuous and therefore a path in the sense of Definition 2.4.13. Also, the track of a polygonal path (or p-path) γ is a union of line segments: $\gamma^* = \cup_{j=1}^{k} [x^{(j-1)}, x^{(j)}]$.

Definition 2.4.16 A subset E of a metric space X is called **path-connected** if, given any $x, y \in E$, there is a path in E with initial point x and terminal point y. If $X = \mathbf{R}^n$, E is said to be **polygonally connected** if, given any $x, y \in E$, there is a polygonal path in E which joins x to y.

Example 2.4.17

(i) Let $a \in \mathbf{R}$. Then $\mathbf{R}\backslash\{a\}$, with the metric inherited from $\mathbf{R}$, is not path-connected. For let $x, y \in \mathbf{R}$, with $x < a < y$, and suppose there is a path $\gamma : [0, 1] \to \mathbf{R}\backslash\{a\}$ joining x to y. By the intermediate-value theorem, $\gamma(t) = a$ for some $t \in [0, 1]$, and we have a contradiction.

(ii) Let $\mathbf{a} \in \mathbf{R}^2$. Then $\mathbf{R}^2\backslash\{\mathbf{a}\}$, with the metric inherited from $\mathbf{R}^2$, is path-connected. For let $\mathbf{x}, \mathbf{y} \in \mathbf{R}^2\backslash\{\mathbf{a}\}$, $\mathbf{x} \neq \mathbf{y}$. Then, if $\mathbf{x}, \mathbf{y}$ and $\mathbf{a}$ are not collinear, the line segment joining $\mathbf{x}$ to $\mathbf{y}$ is a path in $\mathbf{R}^2\backslash\{\mathbf{a}\}$; and if these three points are collinear, $\mathbf{x}$ may be joined to $\mathbf{y}$ by a $p-$path in $\mathbf{R}^2\backslash\{\mathbf{a}\}$ whose track is a union of at most three line segments, each parallel to one of the coordinate axes. The same argument shows that when $n > 2$, removal of one point from $\mathbf{R}^n$ leaves the set path-connected.

Proposition 2.4.18 *Let X and Y be homeomorphic metric spaces. Then X is path-connected if, and only if, Y is path-connected.*

Proof Let $\phi : X \to Y$ be a homeomorphism, suppose that X is path-connected and let $y_1, y_2 \in Y$. Then $y_1 = \phi(x_1)$, $y_2 = \phi(x_2)$ for some $x_1, x_2 \in X$; let $\gamma : [0, 1] \to X$ be a path joining x_1 to x_2. Then $\phi \circ \gamma$ is a path in Y joining y_1 to y_2, and so Y is path-connected. The result is now clear. □

Corollary 2.4.19 *If $n > 1$, $\mathbf{R}$ and $\mathbf{R}^n$ are not homeomorphic.*

Proof Suppose the result is false Then for some $n > 1$, there is a homeomorphism $\phi : \mathbf{R} \to \mathbf{R}^n$. Let $a \in \mathbf{R}$: then the restriction of ϕ to $\mathbf{R}\backslash\{a\}$ is a homeomorphism of $\mathbf{R}\backslash\{a\}$ onto $\mathbf{R}^n\backslash\{\phi(a)\}$. But by Example 2.4.17 (i) and (ii), $\mathbf{R}\backslash\{a\}$ is not path-connected while $\mathbf{R}^n\backslash\{\phi(a)\}$ is path-connected. This contradicts Proposition 2.4.18 and completes the proof. □

We remark that it is also true that if $m, n \in \mathbf{N}$ and $m \neq n$, then $\mathbf{R}^m$ and $\mathbf{R}^n$ are not homeomorphic. However, this is much harder to prove.

Next we relate the notions of connectedness and path-connectedness.

Theorem 2.4.20 *Let E be a path-connected subset of a metric space X. Then E is a connected set in X.*

Proof Suppose E is not connected. Then there are disjoint open sets U, V in X such that

$$U \cap E \neq \emptyset,\ \ V \cap E \neq \emptyset \text{ and } E \subset U \cup V.$$

Let $x \in U \cap E$, $y \in V \cap E$; as E is path-connected, there is a path $\gamma : [0, 1] \to E$ with initial point x and terminal point y. Since γ is continuous, $\gamma^{-1}(U)$ and $\gamma^{-1}(V)$ are open sets in $[0, 1]$; also $\gamma^{-1}(U) \cup \gamma^{-1}(V) = [0, 1]$ and $\gamma^{-1}(U) \cap \gamma^{-1}(V) = \emptyset$. Thus $[0, 1]$ is not connected, contradicting Theorem 2.4.8. □

Example 2.4.21

(i) Every open ball in $\mathbf{R}^n$ $(n \geq 1)$ is connected, as is $\mathbf{R}^n$ itself.
To see this, let $\mathbf{a} \in \mathbf{R}^n$ and $r > 0$. We show that $B(\mathbf{a}, r)$ is path-connected and therefore connected. Let d denote the Euclidean metric on $\mathbf{R}^n$. Let $\mathbf{x}, \mathbf{y} \in B(\mathbf{a}, r)$ and let $\gamma : [0, 1] \to \mathbf{R}^n$ be given by $\gamma(t) = (1 - t)\mathbf{x} + t\mathbf{y}$. We claim that γ is a path in $B(\mathbf{a}, r)$ joining $\mathbf{x}$ to $\mathbf{y}$. Evidently γ is continuous: if $t_n \in [0, 1]$ $(n \in \mathbf{N})$

and $t_n \to t \in [0, 1]$, then $d(\gamma(t_n), \gamma(t)) = |t_n - t|\, d(x, y) \to 0$. Moreover, $\gamma^* \subset B(\mathbf{a}, r)$: for all $t \in [0, 1]$,

$$\gamma(t) - \mathbf{a} = (1 - t)(\mathbf{x} - \mathbf{a}) + t(\mathbf{y} - \mathbf{a})$$

and

$$d(\gamma(t), \mathbf{a}) = \left(\sum_{j=1}^{n} \left\{ (1 - t)(x_j - a_j) + t(y_j - a_j) \right\}^2 \right)^{1/2}$$

$$\leq (1 - t)d(\mathbf{x}, \mathbf{a}) + td(\mathbf{y}, \mathbf{a}) < r.$$

The rest is clear.

(ii) The converse of Theorem 2.4.20 is false: not every connected set is path-connected. To illustrate this, take $X = \mathbf{R}^2$ and

$$E = \{(0, y) : -1 \leq y \leq 1\} \cup \left\{ \left(x, \sin \frac{\pi}{x} \right) : 0 < x \leq 1 \right\} = A \cup B, \text{ say.}$$

The set B is the image of $(0, 1]$ under the continuous map $t \longmapsto \left(t, \sin \frac{\pi}{t}\right)$ and so is connected. We claim that $B \subset A \cup B \subset \overline{B}$: granted this, it follows from Exercise 2.4.33/1 that E is connected. To establish our claim, let $(0, y) \in A$ and let $\varepsilon > 0$; let $n \in \mathbf{N}$ be so large that $1/n < \varepsilon$. Since $\sin\left(2n \pm \frac{1}{2}\right)\pi = \pm 1$, there exists $t \in \left[\frac{2}{4n+1}, \frac{2}{4n-1}\right]$ such that $\sin \frac{\pi}{t} = y$. The point $\left(t, \sin \frac{\pi}{t}\right)$ belongs to B and its distance from $(0, y)$ is less than ε; thus $A \cup B \subset \overline{B}$ and the claim is justified.

However, E is not path-connected. For suppose $\gamma : [0, 1] \to E$ is a path in E with initial and terminal points $(0, 0)$ and $(1, 0)$, respectively; write $\gamma(t) = (\gamma_1(t), \gamma_2(t))$, $t \in [0, 1]$. Then $\gamma^{-1}(A)$ is a closed set contained in $[0, 1]$ and containing 0; thus $b := \sup \gamma^{-1}(A) \in \gamma^{-1}(A)$ and $0 \leq b < 1$. Suppose that $\gamma_2(b) \leq 0$. Then given any $\delta > 0$ with $b + \delta \leq 1$, we have $\gamma_1(b + \delta) > 0$, and there exists $n \in \mathbf{N}$ such that

$$0 = \gamma_1(b) < 2/(4n + 1) < \gamma_1(b + \delta);$$

also, by the intermediate-value theorem, there exists t such that $b < t < b + \delta$ and $\gamma_1(t) = 2/(4n + 1)$. Hence $\gamma_2(t) = 1$ and $\gamma_2(t) - \gamma_2(b) \geq 1$. The same kind of argument may be used if $\gamma_2(b) \geq 0$, and we conclude that γ_2 is not continuous at b. This contradiction shows that E is not path-connected.

(iii) The closure of a path-connected set need not be path-connected. For with the notation of (ii), B is plainly path-connected but E, and hence $\overline{B}$, are not.

(iv) For any $n \in \mathbf{N}$, the *unit sphere* S^n in $\mathbf{R}^{n+1}$ is a connected subset of $\mathbf{R}^{n+1}$. To see this, note that by Example 2.4.17 (ii), $\mathbf{R}^{n+1}\backslash\{0\}$ is path-connected; by Theorem 2.4.20 it is connected. Define $f : \mathbf{R}^{n+1}\backslash\{0\} \to S^n$ by

$$f(x_1, \ldots, x_n) = (x_1, \ldots, x_n)/(x_1^2 + \ldots + x_n^2)^{1/2}.$$

Since f is clearly continuous and surjective, it follows from Theorem 2.4.11 that S^n is connected.

In view of Example 2.4.21 (ii) above, it is a relief to know that provided that we restrict ourselves to open subsets of $\mathbf{R}^n$, the notions of connectedness and path-connectedness coincide. The next lemma prepares for this result.

Lemma 2.4.22 *Let $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in \mathbf{R}^n$. Then there is a map $\gamma : [0, 1] \to \mathbf{R}^n$ which is a p-path in $\mathbf{R}^n$ joining x to y such that*

$$d(\gamma(s), \gamma(t)) \leq d(x, y)\ (s, t \in [0, 1])$$

where d is the Euclidean metric on $\mathbf{R}^n$.

Proof Let $e^{(1)}, \ldots, e^{(n)}$ be the vectors of the natural basis for $\mathbf{R}^n$. Let $p^{(0)} = x$ and

$$p^{(j)} = x + \sum_{k=1}^{j} (y_k - x_k)e^{(k)}\ (j = 1, \ldots, n).$$

Define $\gamma : [0, 1] \to \mathbf{R}^n$ by

$$\gamma(s) = (j - ns)p^{(j-1)} + (ns - j + 1)p^{(j)}$$

if $j - 1 \leq ns \leq j$ and j is a positive integer not exceeding n. It is routine to verify that, for all $s \in [0, 1]$,

$$\gamma(s) = x + \sum_{k=1}^{n} \Psi_k(s)(y_k - x_k)e^{(k)},$$

where

$$\Psi_k(s) = \min\{\max\{ns - k + 1, 0\}, 1\}.$$

Hence γ is a p-path in $\mathbf{R}^n$ joining x to y; moreover, since $0 \leq \Psi_k(s), \Psi_k(t) \leq 1$ and therefore $|\Psi_k(s) - \Psi_k(t)| \leq 1$, we have for all $s, t \in [0, 1]$,

$$d(\gamma(s), \gamma(t)) = \left\{ \sum_{k=1}^{n} |\Psi_k(s) - \Psi_k(t)|^2\, |y_k - x_k|^2 \right\}^{1/2} \leq d(x, y).$$

□

Theorem 2.4.22 *Let G be an open set in $\mathbf{R}^n$. Then the following statements are equivalent.*

(i) *G is connected.*
(ii) *G is polygonally connected; further, given any $x, y \in G$ there is a p-path in G joining them.*
(iii) *G is path-connected.*

Proof Suppose that $G \neq \emptyset$; otherwise, the result holds trivially. It is obvious that (ii) implies (iii); also, Theorem 2.4.20 shows that (iii) implies (i). It remains to prove that (i) implies (ii).

Suppose that G is connected, let $a \in G$ and let

$$A := \{x \in G : \text{there is a p-path in } G \text{ joining } a \text{ to } x\}.$$

To show that G is polygonally connected it is enough to prove that $A = G$. First we prove that A is open. Let $x \in A$ and let $\mu : [0, 1] \to \mathbf{R}^n$ be a p-path in G joining a to x. Since $x \in G$, there exists $r > 0$ such that $B(x, r) \subset G$. Let $y \in B(x, r)$. By Lemma 2.4.22, there is a map $\nu : [0, 1] \to \mathbf{R}^n$ which is a p-path in $B(x, r)$, and hence in G, joining x to y. Let $\gamma : [0, 1] \to \mathbf{R}^n$ be defined by

$$\gamma(t) = \begin{cases} \mu(2t) & \text{if } 0 \leq t \leq \frac{1}{2}, \\ \nu(2t-1) & \text{if } \frac{1}{2} \leq t \leq 1. \end{cases}$$

Evidently γ is a p-path in G joining a to y. Hence $y \in A$. It follows that $B(x, r) \subset A$ and that A is open.

Next we show that $G \backslash A$ is open. Let $z \in G \backslash A$ and let $r' > 0$ be such that $B(z, r') \subset G$. It is enough to prove that $B(z, r') \subset G \backslash A$. To obtain a contradiction, suppose that this is not the case. Then there exist $w \in B(z, r') \cap A$ and a p-path in G joining a to w. Further, this path may be extended, by means of a construction similar to that of the previous paragraph, to a p-path in G joining a to z. It follows that $z \in A \cap (G \backslash A)$, an impossibility.

Finally, note that $a \in A$, $G = A \cup (G \backslash A)$ and $A \cap (G \backslash A) = \emptyset$. Thus, since G is connected, $G \backslash A = \emptyset$ and $A = G$. □

Next we turn to components: the idea is that even if a set is not connected, it is made up of connected subsets; components are the largest such subsets.

Definition 2.4.24 Let E be a non-empty subset of a metric space X. A subset D of E is called a **component** of E if it is a maximal connected subset of E, that is, if (i) D is connected, and (ii) whenever D_1 is connected and $D \subset D_1 \subset E$, it follows that $D = D_1$.

To prove the basic theorem about components, the following lemma will be very useful.

Lemma 2.4.25 *Let E be a non-empty subset of a metric space X and let $\mathscr{F}$ be a non-empty family of connected subsets of E with one point in common; that is, there exists $a \in \cap\, \mathscr{F}$. Then $A := \cup\mathscr{F}$ is a connected subset of E.*

Proof In view of Lemma 2.1.5 (iii) it is enough to show that A is a connected subset of X. Suppose that this is not so. Then there are disjoint open sets U, V in X such that $A \cap U \neq \emptyset$, $A \cap V \neq \emptyset$ and $A \subset U \cup V$. Since each $F \in \mathscr{F}$ is connected and $F \subset U \cup V$, either $F \cap U = \emptyset$ or $F \cap V = \emptyset$. As $A \cap U \neq \emptyset$, there exists $F_1 \in \mathscr{F}$ such that $F_1 \cap U \neq \emptyset$ and so $F_1 \cap V = \emptyset$. Since $A \cap V \neq \emptyset$, there exists $F_2 \in \mathscr{F}$ such that $F_2 \cap V \neq \emptyset$ and so $F_2 \cap U = \emptyset$. Hence $a \in F_1 \cap F_2 = F_1 \cap F_2 \cap (U \cup V) \subset (F_2 \cap U) \cup (F_1 \cap V) = \emptyset$, and we have a contradiction. □

Theorem 2.4.26 *Let E be a non-empty subset of a metric space X. Then*

(i) *each $a \in E$ lies in a component of E (so that E is the union of its components);*
(ii) *distinct components of E are disjoint.*

Proof Let $a \in E$ and let $\mathscr{F}$ be the family of all connected subsets of E which contain a. Plainly $\mathscr{F} \neq \emptyset$, since $\{a\} \in \mathscr{F}$. By Lemma 2.4.25, $A := \cup\mathscr{F}$ is connected and contains a. Now A is a component of E : for, if $A \subset A_1 \subset E$ and A_1 is connected, then $A_1 \in \mathscr{F}$ and so $A_1 = A$. This proves (i).

Regarding (ii), let A_1 and A_2 be components of E, suppose that $A_1 \neq A_2$ and that $a \in A_1 \cap A_2$. By Lemma 2.4.25 , $A_1 \cup A_2$ is connected. But in that event, since A_1 and A_2 are components, it follows that $A_1 = A_1 \cup A_2 = A_2$, a contradiction. □

Theorem 2.4.27 *Let G be a non-empty open subset of $\mathbf{R}^n$. Then G has countably many components, each of which is open.*

Proof Let A be a component of G and let $a \in A$. Since G is open, there exists $\varepsilon > 0$ such that $B(a, \varepsilon) \subset G$. Now $B(a, \varepsilon)$ is path-connected and thus connected: hence, by Lemma 2.4.25, $A \cup B(a, \varepsilon)$ is connected. As A is a component this implies that $A \cup B(a, \varepsilon) = A$. Hence $B(a, \varepsilon) \subset A$ and A is open.

The set $\mathbf{Q}^n$ is a countable subset of $\mathbf{R}^n$ and may be written as $\{p_k : k \in \mathbf{N}\}$. Given any component A of G, there exists a least $k \in \mathbf{N}$ such that $p_k \in A$. By Theorem 2.4.26, to distinct components there correspond distinct k, and so the components may be put in one-to-one correspondence with a subset of $\mathbf{N}$. The proof is complete. □

Corollary 2.4.28 *Let $G \subset \mathbf{R}$ be open. Then $G = \bigcup_{n=1}^{\infty} I_n$, where the I_n are pairwise disjoint open intervals.*

Companion to the notion of a component of a set there is that of a path-component.

Definition 2.4.29 A **path-component** of a subset A of a metric space X is a maximal path-connected subset of A.

This idea has useful consequences, given below. Note that, plainly, distinct path-components are disjoint.

Theorem 2.4.30 *Each path-component of a metric space X is open (and therefore also closed) if, and only if, each point of X has a path-connected neighbourhood. The space X is path-connected if, and only if, it is connected and each $x \in X$ has a path-connected neighbourhood.*

Proof Suppose that each path-component of X is open, and let $x \in X$. Let C be the path-component containing x : C is a neighbourhood of x and is path-connected. Conversely, suppose that each point of X has a path-connected neighbourhood, let C be a path-component and let $x \in C$. Then there is a path-connected neighbourhood $U(x)$ of x, and since C is a maximal path-connected set containing x, $U(x) \subset C$. Thus $C = \bigcup_{x \in C} U(x)$ is open. Since $X \backslash C$ is the union of the remaining open path-components, it is open: thus C is closed.

If X is path-connected it is connected, by Theorem 2.4.20, and, of course, X is a path-connected neighbourhood of its points. Conversely, suppose that X is connected and that each $x \in X$ has a path-connected neighbourhood. Then each path-component is both open and closed; and since X is connected, this path-component must be X. $\square$

To conclude this section we show that connectedness and path-connectedness are preserved on taking products.

Theorem 2.4.31 *Let X_1, X_2 be connected (respectively, path-connected) metric spaces. Then the metric space $X_1 \times X_2$ (see Example 2.1.2 (ix)) is connected (respectively, path-connected).*

Proof First suppose that X_1 and X_2 are connected and let $(a_1, a_2), (b_1, b_2) \in X_1 \times X_2$. Then $\{a_1\} \times X_2$ and $X_1 \times \{b_2\}$ are connected subsets of $X_1 \times X_2$ as they are homeomorphic (even isometric) to X_2 and X_1 respectively; moreover, they have a common point, (a_1, b_2). By Lemma 2.4.25 their union is connected: thus there is a connected set containing (a_1, a_2) and (b_1, b_2). The connectedness of $X_1 \times X_2$ now follows from Example 2.4.7 (v).

Now suppose that X_1 and X_2 are path-connected and again let $(a_1, a_2), (b_1, b_2) \in X_1 \times X_2$. There is a path $\gamma_1 : [0, 1] \to X_1$ joining a_1 to b_1, and hence there is a path $\widetilde{\gamma}_1 : [0, 1] \to X_1 \times X_2$ joining (a_1, b_2) to (b_1, b_2), given by $\widetilde{\gamma}_1(t) = (\gamma_1(t), b_2)$. Similarly, there is a path $\widetilde{\gamma}_2 : [0, 1] \to X_1 \times X_2$ joining (a_1, a_2) to (a_1, b_2). The path $\widetilde{\gamma} : [0, 1] \to X_1 \times X_2$ defined by

$$\widetilde{\gamma}(t) = \begin{cases} \widetilde{\gamma}_2(2t) & \text{if } 0 \le t \le \frac{1}{2}, \\ \widetilde{\gamma}_1(2t - 1) & \text{if } \frac{1}{2} \le t \le 1 \end{cases}$$

joins (a_1, a_2) to (b_1, b_2), and so $X_1 \times X_2$ is path-connected. $\square$

Corollary 2.4.32 *The torus $T := S^1 \times S^1$ is connected.*

Proof From Example 2.4.21 (iv) we see that S^1 is connected. The corollary now follows from Theorem 2.4.31. $\square$

Of course, the torus as defined here is a subset of $\mathbf{R}^4$ and is endowed with the inherited metric. In fact, T is homeomorphic to the subset $\widetilde{T}$ of $\mathbf{R}^3$ obtained by revolution of the circle $\{(0, y, z) : (y-1)^2 + z^2 = 1/4\}$ about the z-axis. For $\widetilde{T}$ is given parametrically by

$$x = \left(1 + \frac{\cos\theta}{2}\right)\cos\phi,\ \ y = \left(1 + \frac{\cos\theta}{2}\right)\sin\phi,$$

$$z = \frac{\sin\theta}{2}(0 \le \theta < 2\pi,\ \ 0 \le \phi < 2\pi),$$

and the map

$$((\cos\theta, \sin\theta), (\cos\phi, \sin\phi)) \longmapsto \left(\left(1 + \frac{\cos\theta}{2}\right)\cos\phi, \left(1 + \frac{\cos\theta}{2}\right)\sin\phi, \frac{\sin\theta}{2}\right)$$

is a homeomorphism of T onto $\widetilde{T}$. This map, f, is given by

$$f((a, b), (c, d)) = \left(\left(1 + \frac{1}{2}a\right)c, \left(1 + \frac{1}{2}a\right)d, \frac{1}{2}b\right)$$

and is evidently continuous. It is bijective, with

$$f^{-1}(p, q, r) = \left(2\left(-1 + \sqrt{p^2 + q^2}\right), 2r, \frac{p}{\sqrt{p^2 + q^2}}, \frac{q}{\sqrt{p^2 + q^2}}\right)$$

since $p = \left(1 + \frac{1}{2}a\right)c$, $q = \left(1 + \frac{1}{2}a\right)d$, $r = \frac{1}{2}b$, and so

$$p^2 + q^2 = \left(1 + \frac{1}{2}a\right)^2,\ \ \frac{1}{2}a = -1 + \sqrt{p^2 + q^2}$$

since $1 + \frac{1}{2}a \ge \frac{1}{2}$. Plainly f^{-1} is continuous, and so f is a homeomorphism.

Exercise 2.4.33

1. Let A be a connected subset of a metric space and suppose that $A \subset B \subset \overline{A}$. Prove that B is connected. Deduce that the components of a closed set are closed.
2. Let $\mathbf{R}^2$ be endowed with the Euclidean metric and let S be a subset of $\mathbf{R}^2$ which is both open and closed. Prove that either $S = \emptyset$ or $S = \mathbf{R}^2$.
3. Let E and F be subsets of $\mathbf{R}^2$ (endowed with the Euclidean metric) defined by

$$E = \{(x, y) : x^2 + y^2 \le 1\} \cup \{(x, y) : (x-2)^2 + y^2 < 1\},$$
$$F = \{(x, y) : x^2 + y^2 < 1\} \cup \{(1 + 1/n, 0) : n \in \mathbf{N}\}.$$

Determine whether E or F is connected. What are the components of these sets?

4. Let $n \in \mathbf{N}$ and let $GL(n, \mathbf{R})$ be the set of all non-singular $n \times n$ matrices; identify $GL(n, \mathbf{R})$ with a subset of $\mathbf{R}^{n^2}$ in an obvious way and give it the inherited metric. Prove that $GL(n, \mathbf{R})$ is not connected.
5. Let A and B be path-connected subsets of a metric space such that $A \cap B \neq \emptyset$. Prove that $A \cup B$ is path-connected.
6. Let E and F be metric spaces, with E path-connected, and let $f : E \to F$ be continuous. Prove that $f(E)$ is path-connected.
7. Let K be the subset of $[0, 1]$ consisting of all numbers of the form $\sum_{n=0}^{\infty} 3^{-n} c_n$, with $c_n \in \{0, 2\}$ for all $n \in \mathbf{N}_0$. This set is called the **Cantor** set (see Exercise 2.3.38/16). Show that K is compact, that $[0, 1]\backslash K$ is a countable union of disjoint intervals, and that the sum of the lengths of these intervals is 1. Show that given any $x \in K$, the connected component of K which contains x is $\{x\}$.
8. Let $S = [0, 1] \times [0, 1]$, let K be as in the question above and let $f : K \to S$ be the map which to each $x \in K$, with $x = \sum_{n=0}^{\infty} 3^{-n} c_n$, assigns the element $\left(\sum_{n=0}^{\infty} 2^{-n} b_{2n+1}, \sum_{n=0}^{\infty} 2^{-n} b_{2n}\right)$, where $b_m = c_m/2 \quad (m \in \mathbf{N}_0)$. Show that f is well-defined, and that it is surjective and continuous. Deduce that there is a continuous surjective map $g : [0, 1] \to S$. (This is Peano's **space-filling** curve.)

2.5 Simple-Connectedness

Our interest here is in those path-connected metric spaces which, loosely speaking, may be viewed as without holes. To bring precision to this, the notion of homotopy is introduced. Throughout this section the closed interval $[0, 1]$ will be denoted by I; and if X is a metric space the product $X \times I$ is assumed to be equipped with the metric of Example 2.1.2 (ix).

Definition 2.5.1 Let X and Y be metric spaces and let $f_0, f_1 : X \to Y$ be continuous. We say that the maps f_0 and f_1 are **homotopic**, and write $f_0 \simeq f_1$, if there is a continuous map $F : X \times I \to Y$ such that, for all $x \in X$,

$$F(x, 0) = f_0(x) \textit{ and } F(x, 1) = f_1(x).$$

Such a map F is called a **homotopy** between f_0 and f_1.

Example 2.5.2 Let X be a metric space and let $f_0, f_1 : X \to \mathbf{R}^n$ be continuous. Define $F : X \times I \to \mathbf{R}^n$ by

$$F(x, t) = (1 - t)f_0(x) + tf_1(x), \;\; (x, t) \in X \times I.$$

Then it is easy to verify that F is a homotopy between f_0 and f_1.

With regard to the homotopy F in Definition 2.5.1, if we set $f_t(x) = F(x, t)$, then $\{f_t : t \in I\}$ is a one-parameter family of continuous maps from X to Y, and we may think of the homotopy as a continuous deformation of f_0 into f_1.

Definition 2.5.3 Let X and Y be metric spaces and let A be a subset of X. Let $f_0, f_1 : X \to Y$ be continuous maps such that $f_0(a) = f_1(a)$ for all $a \in A$, that is, $f_0 \mid_A = f_1 \mid_A$. If there is a homotopy F between f_0 and f_1 such that, for all $a \in A$ and all $t \in I$,

$$F(a, t) = f_0(a) = f_1(a),$$

or equivalently $f_t \mid_A = f_1 \mid_A$ for all $t \in I$, then we say that f_0 and f_1 are **homotopic relative to** A and write $f_0 \simeq f_1$ rel A. [Note: if A is empty, then $\simeq$ rel A and $\simeq$ coincide.]

Example 2.5.4 Let $A = \{0, 1\}$. Let $f_0, f_1 : I \to \mathbf{R}^n$ be paths in $\mathbf{R}^n$ such that $f_0(0) = f_1(0)$ and $f_0(1) = f_1(1)$: the paths have a common initial point and a common terminal point so that $f_0 \mid_A = f_1 \mid_A$. Consideration of $F : I \times I \to \mathbf{R}^n$ given by

$$F(s, t) = (1 - t)f_0(s) + tf_1(s)$$

shows that $f_0 \simeq f_1$ rel $\{0, 1\}$.

Theorem 2.5.5 *Let X and Y be metric spaces and A be a subset of X. Then $\simeq$ rel A is an equivalence relation in $C(X, Y)$, the family of continuous maps from X to Y.*

Proof The steps which follow show that $\simeq$ rel A is reflexive, symmetric and transitive.

(1) If $f \in C(X, Y)$, then $f \simeq f$ rel A.
The continuous map $F : X \times I \to Y$ given by $F(x, t) = f(x)$ verifies this claim.
(2) If $f, g \in C(X, Y)$ and $f \simeq g$ rel A, then $g \simeq f$ rel A.

By hypothesis, there exists a homotopy F relative to A between f and g. Let $G : X \times I \to Y$ be defined by

$$G(x, t) = F(x, 1 - t).$$

As it is a composition of continuous maps, G is continuous. Moreover, for all $x \in X$,

$$G(x, 0) = g(x), \;\; G(x, 1) = f(x);$$

also, for all $a \in A$ and $t \in I$,

$$G(a, t) = g(a) = f(a).$$

Hence $g \simeq f$ rel A.
(3) If $f, g, h \in C(X, Y)$, $f \simeq g$ rel A and $g \simeq h$ rel A, then $f \simeq h$ rel A.

Given that there are homotopies F and G relative to A between f and g, and g and h, respectively, let $H : X \times I \to Y$ be defined by

$$H(x, t) = \begin{cases} F(x, 2t), & 0 \leq t \leq 1/2, \\ G(x, 2t - 1), & 1/2 \leq t \leq 1. \end{cases}$$

Since $F(x, 1) = g(x) = G(x, 0)$ for all $x \in X$, there is consistency of definition on $X \times \{1/2\}$ and, by appeal to the glueing lemma (Lemma 2.1.35), it follows that H is continuous. Further, for all $x \in X$,

$$H(x, 0) = f(x), H(x, 1) = h(x);$$

also, for all $a \in A$ and $t \in I$,

$$H(a, t) = f(a) = h(a).$$

Hence $f \simeq h$ rel A. □

Corollary 2.5.6 *Let $f \in C(X, Y)$ and denote by $\langle f\rangle$ the equivalence class associated with f :*

$$\langle f\rangle = \{g \in C(X, Y) : g \simeq f \text{ rel } A\}.$$

The family of equivalence classes $\{\langle f\rangle : f \in C(X, Y)\}$ constitutes a partition of $C(X, Y)$, by which we mean that no equivalence class is empty, their union exhausts $C(X, Y)$ and, for all $f, g \in C(X, Y)$, the classes $\langle f\rangle$ and $\langle g\rangle$ are either disjoint or identical.

Proof We leave this to the reader, noting that it is a special case of a general result concerning equivalence classes: see, for example, [19], p. 50. □

We now show that relative to the composition of functions, homotopy is well-behaved.

Theorem 2.5.7 *Let X, Y and Z be metric spaces and A be a subset of X. Let f_0, $f_1 : X \to Y$ and $g_0, g_1 : Y \to Z$ be continuous maps such that $f_0 \simeq f_1$ rel A and $g_0 \simeq g_1$ rel $f_0(A)$. Then*

$$g_0 \circ f_0 \simeq g_1 \circ f_1 \text{ rel } A.$$

Proof Let $F : X \times I \to Y$ and $G : Y \times I \to Z$ be homotopies establishing that $f_0 \simeq f_1$ rel A and $g_0 \simeq g_1$ rel $f_0(A)$, respectively. The map $g_0 \circ F : X \times I \to Z$ is continuous; also, for all $x \in X$,

$$(g_0 \circ F)(x, 0) = (g_0 \circ f_0)(x),\ (g_0 \circ F)(x, 1) = (g_0 \circ f_1)(x),$$

and, for all $a \in A$ and $t \in I$,

$$(g_0 \circ F)(a, t) = (g_0 \circ f_0)(a) = (g_0 \circ f_1)(a).$$

Hence $g_0 \circ f_0 \simeq g_0 \circ f_1$ rel A. Next, consider the map $H : X \times I \to Z$ defined by $H(x, t) = G(f_1(x), t)$. It is continuous; moreover, for all $x \in X$,

$$H(x, 0) = (g_0 \circ f_1)(x),\ H(x, 1) = (g_1 \circ f_1)(x),$$

and for all $a \in A$ and $t \in I$,

$$H(a, t) = (g_0 \circ f_1)(a) = (g_1 \circ f_1)(a).$$

Thus $g_0 \circ f_1 \simeq g_1 \circ f_1$ rel A. Finally, by Theorem 2.5.5, $g_0 \circ f_0 \simeq g_1 \circ f_1$ rel A. □

In the next section we appeal to the simplest aspect of this theorem, when $g_0 = g_1$.

2.5.1 Homotopies Between Paths

Let X be a metric space. For present purposes we shall think of $C(I, X)$ as the set of all paths in X, each path being assumed to have $I = [0, 1]$ as its parameter interval. For brevity, the symbol $\sim$ will be used for the relation $\simeq$ rel $\{0, 1\}$ on $C(I, X)$. Hence $f_0 \sim f_1$, to be read f_0 is equivalent to f_1, is understood to mean that $f_0(0) = f_1(0)$, $f_0(1) = f_1(1)$ and that a continuous map $F : I \times I \to X$ exists such that

$$F(s, 0) = f_0(s),\ F(s, 1) = f_1(s)\ (s \in I)$$

and

$$F(0, t) = f_0(0),\ F(1, t) = f_0(1)\ (t \in I).$$

The homotopy F may be viewed as continuously deforming f_0 into f_1 through a family of paths with prescribed endpoints.

Definition 2.5.8 Let f and g be paths in a metric space X such that $f(1) = g(0)$. The product path $f * g : I \to X$ is defined by

$$(f * g)(s) = \begin{cases} f(2s), & 0 \le s \le 1/2, \\ g(2s - 1), & 1/2 \le s \le 1. \end{cases}$$

Similarly, if $f_1, f_2, \dots, f_n : I \to X$ are paths in X such that, for $1 \le j \le n - 1$, $f_j(1) = f_{j+1}(0)$, then the product path $f_1 * f_2 * \dots * f_n : I \to X$ is defined by

$$(f_1 * f_2 * \dots * f_n)(s) = \begin{cases} f_1(ns), & 0 \le s \le 1/n, \\ f_2(ns - 1), & 1/n \le s \le 2/n, \\ \dots & \dots \\ f_j(ns - j + 1), & (j - 1)/n \le s \le j/n, \\ \dots & \dots \\ f_n(ns - n + 1), & (n - 1)/n \le s \le 1. \end{cases}$$

Evidently $f * g$ is a path joining $f(0)$ to $g(1)$; likewise, $f_1 * f_2 * \dots * f_n$ is a path joining $f_1(0)$ to $f_n(1)$.

Theorem 2.5.9 *Let f, f', g and g' be paths in a metric space X; suppose that $f \sim f'$, $g \sim g'$ and $f * g$ is defined. Then $f * g \sim f' * g'$.*

Proof Let f, f' join x to y and g, g' join y to z : since $f * g$ is defined, $f(1) = y = g(0)$. As $f \sim f'$, there exists a continuous map $F : I \times I \to X$ such that

$$F(s, 0) = f(s),\ F(s, 1) = f'(s)\ (s \in I)$$

and

$$F(0, t) = x,\ F(1, t) = y\ (t \in I).$$

Similarly, since $g \sim g'$, there is a continuous map $G : I \times I \to X$ such that

$$G(s, 0) = g(s),\ G(s, 1) = g'(s)\ (s \in I)$$

and

$$G(0, t) = y,\ G(1, t) = z\ (t \in I).$$

Let $H : I \times I \to X$ be defined by

$$H(s, t) = \begin{cases} F(2s, t), & 0 \le s \le 1/2,\ 0 \le t \le 1, \\ G(2s - 1, t), & 1/2 \le s \le 1,\ 0 \le t \le 1. \end{cases}$$

Since $F(1, t) = y = G(0, t)$ for all $t \in I$, there is consistency of definition on the line segment $\{1/2\} \times I$. The glueing lemma ensures that H is continuous; further, for all $s \in I$,

$$H(s, 0) = (f * g)(s),\ H(s, 1) = (f' * g')(s)$$

and, for all $t \in I$,

$$H(0, t) = x,\ H(1, t) = z.$$

Hence $f * g \sim f' * g'$. □

By dividing the unit square $I \times I$ into n vertical strips rather than 2, the following generalisation of the last theorem may be established: details are left to the reader.

Theorem 2.5.10 *Let $f_1, f_2, \ldots, f_n$ and $f_1', f_2', \ldots, f_n'$ be paths in a metric space X; suppose that, for $1 \le j \le n$, $f_j \sim f_j'$ and that the product path $f_1 * f_2 * \ldots * f_n$ is defined. Then*

$$f_1 * f_2 * \ldots * f_n \sim f_1' * f_2' * \ldots * f_n'.$$

Theorem 2.5.11 *Let $f_1, f_2, \ldots, f_n$ $(n \ge 3)$ be paths in a metric space X such that the product $f_1 * f_2 * \ldots * f_n$ is defined. Suppose $1 \le k \le n - 1$ and let the map $\phi : I \to I$ be given by*

$$\phi(s) = \begin{cases} 2sk/n, & 0 \le s \le 1/2, \\ (n - k)(2s - 1)/n + k/n, & 1/2 \le s \le 1. \end{cases}$$

Then

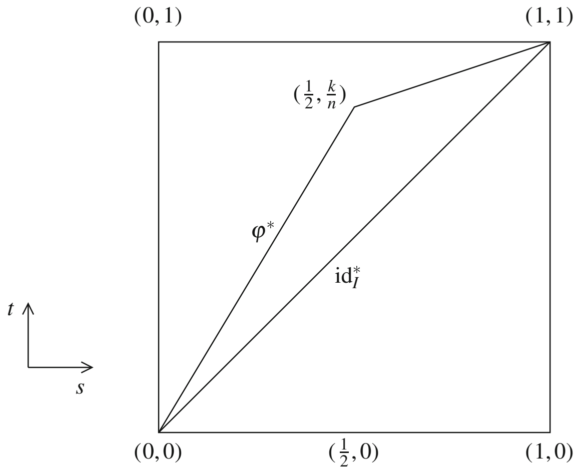
(i) $(f_1 * f_2 * \ldots * f_n) \circ \phi = (f_1 * f_2 * \ldots * f_k) * (f_{k+1} * \ldots * f_n)$;
(ii) $(f_1 * f_2 * \ldots * f_k) * (f_{k+1} * \ldots * f_n) \sim f_1 * f_2 * \ldots * f_n$;

and

(iii) *setting* $n = 3$, $(f_1 * f_2) * f_3 \sim f_1 * (f_2 * f_3)$.

Proof

(i) Illustrated below, the map ϕ is continuous and strictly increasing; $\phi(s) \leq k/n$ if $0 \leq s \leq 1/2$, $\phi(s) > k/n$ if $1/2 < s \leq 1$.



Hence, for all $s \in I$,

$$
\begin{aligned}
&((f_1 * f_2 * \ldots * f_n) \circ \phi)(s) \\
&\quad = f_j(n\phi(s) - j + 1) \quad \text{if} \quad (j-1)/n \leq \phi(s) \leq j/n \quad \text{and} \quad 1 \leq j \leq n \\
&\quad = \begin{cases} f_j(2ks - j + 1), & j - 1 \leq 2sk \leq j,\ 1 \leq j \leq k, \\ f_j((n-k)(2s-1) + k - j + 1), & j - k - 1 \leq (2s-1)(n-k) \leq j - k, \\ & k + 1 \leq j \leq n, \end{cases} \\
&\quad = \begin{cases} f_j(2ks - j + 1), & j - 1 \leq 2sk \leq j,\ 1 \leq j \leq k, \\ f_{k+j'}((n-k)(2s-1) - j' + 1), & j' - 1 \leq (2s-1)(n-k) \leq j', \\ & 1 \leq j' \leq n - k, \end{cases} \\
&\quad = ((f_1 * f_2 * \ldots * f_k) * (f_{k+1} * \ldots * f_n))(s).
\end{aligned}
$$

(ii) Consideration of the map $H : I \times I \to I$ given by

$$H(s, t) = (1 - t)\phi(s) + ts$$

shows that ϕ and id_I (the identity map on I) are homotopic relative to $\{0, 1\}$. Hence, using Theorem 2.5.7,

$$\begin{aligned}(f_1 * f_2 * \ldots * f_k) * (f_{k+1} * \ldots * f_n) &= (f_1 * f_2 * \ldots * f_n) \circ \phi \\ &\sim (f_1 * f_2 * \ldots * f_n) \circ id_I \\ &= f_1 * f_2 * \ldots * f_n.\end{aligned}$$

(ii) By (ii), both the product paths $f_1 * (f_2 * f_3)$ and $(f_1 * f_2) * f_3$ are equivalent to $f_1 * f_2 * f_3$. Since the relation $\sim$ is transitive, it follows that $f_1 * (f_2 * f_3) \sim (f_1 * f_2) * f_3$. □

Theorem 2.5.12 *Let X be a metric space, let $x, y \in X$ and let e_x, e_y be the constant paths in X defined by $e_x(s) = x$ and $e_y(s) = y$ $(s \in I)$. Let f be a path in X such that $f(0) = x$ and $f(1) = y$. Then $e_x * f \sim f$ and $f * e_y \sim f$.*

Proof As each equivalence has a similar proof we give only that which involves $e_x * f$. Let $\psi : I \to I$ be given by

$$\psi(s) = \begin{cases} 0, & 0 \le s \le 1/2, \\ 2s - 1, & 1/2 \le s \le 1. \end{cases}$$

The continuous map $H : I \times I \to I$ defined by $H(s, t) = (1 - t)\psi(s) + ts$ enables us to see that ψ and id_I (the identity map on I) are homotopic relative to $\{0, 1\}$. Hence, noting that for all $s \in I$,

$$\begin{aligned}(f \circ \psi)(s) &= \begin{cases} x, & 0 \le s \le 1/2, \\ f(2s - 1), & 1/2 \le s \le 1, \end{cases} \\ &= (e_x * f)(s),\end{aligned}$$

application of Theorem 2.5.7 shows that

$$e_x * f = f \circ \psi \sim f \circ id_I = f,$$

as required. □

Theorem 2.5.13 *Let X be a metric space, f be a path in X and $\widehat{f}$ be the path defined by $\widehat{f}(s) = f(1 - s)$ $(s \in I)$; $\widehat{f}$ is termed the **reverse** of f. Let $f(0) = x$ and $f(1) = y$. Then*

$$f * \widehat{f} \sim e_x\,, \ \widehat{f} * f \sim e_y,$$

where e_x, e_y are the constant paths given by $e_x(s) = x$, $e_y(s) = y$ $(s \in I)$, respectively.

Proof Since the rôles of f and $\widehat{f}$ can be interchanged, it is sufficient to prove that $f * \widehat{f} \sim e_x$. Let $\tau, \theta : I \to I$ be given by

$$\tau(s) = \begin{cases} 2s, & 0 \le s \le 1/2, \\ 2(1 - s), & 1/2 \le s \le 1, \end{cases}$$

and $\theta(s) = 0$ $(s \in I)$. The map

$$(s, t) \mapsto (1 - t)\tau(s) : I \times I \to I$$

shows that $\tau \simeq \theta$ rel $\{0, 1\}$. Since

$$\begin{aligned}(f * \widehat{f})(s) &= \begin{cases} f(2s), & 0 \le s \le 1/2, \\ \widehat{f}(2s-1), & 1/2 \le s \le 1 \end{cases} \\ &= \begin{cases} f(2s), & 0 \le s \le 1/2, \\ f(2(1-s)), & 1/2 \le s \le 1, \end{cases} \\ &= (f \circ \tau)(s),\end{aligned}$$

so that $f * \widehat{f} = f \circ \tau$, application of Theorem 2.5.7 shows that $f * \widehat{f} = f \circ \tau \sim f \circ \theta = e_x$. □

Definition 2.5.14 A **closed path** (or **loop**) in a metric space X is a path whose initial and terminal points coincide: this common point is called its **base point** . Thus, if $x \in X$ and f is a path in X such that $f(0) = f(1) = x$, then f is a closed path in X with base point x.

Remark 2.5.15

(i) Each $x \in X$ is a base point for at least one closed path in X, namely e_x, given by $e_x(s) = x$ $(s \in I)$, the path constant at x.
(ii) If $x, y \in X$ and there is a path f joining x to y then, with $\widehat{f}$ denoting the path given by $\widehat{f}(s) = f(1 - s)$, $f * \widehat{f}$ is a closed path with base point x and $\widehat{f} * f$ is a closed path with base point y.

The definition to follow introduces a new type of homotopy, specific to closed paths, called **free homotopy**. For closed paths f, g in a metric space X recall that the statement $f \simeq g$ rel $\{0, 1\}$, more simply denoted $f \sim g$, means that f and g have a common base point and that a continuous map $H : I \times I \to X$ exists such that

(1) $H(s, 0) = f(s)$, $H(s, 1) = g(s)$ $(s \in I)$

and

(2) $H(0, t) = f(0) = H(1, t)$ $(t \in I)$.

Note that the one-parameter family $\{h_t\}$ of paths determined by H is made up of closed paths with a common base point. The notion of free homotopy relaxes condition (2).

Definition 2.5.16 Let f and g be closed paths in a metric space X. Then f is said to be **freely homotopic** to g if there is a continuous map $H : I \times I \to X$ such that

(i) $H(s, 0) = f(s)$, $H(s, 1) = g(s)$ $(s \in I)$

and

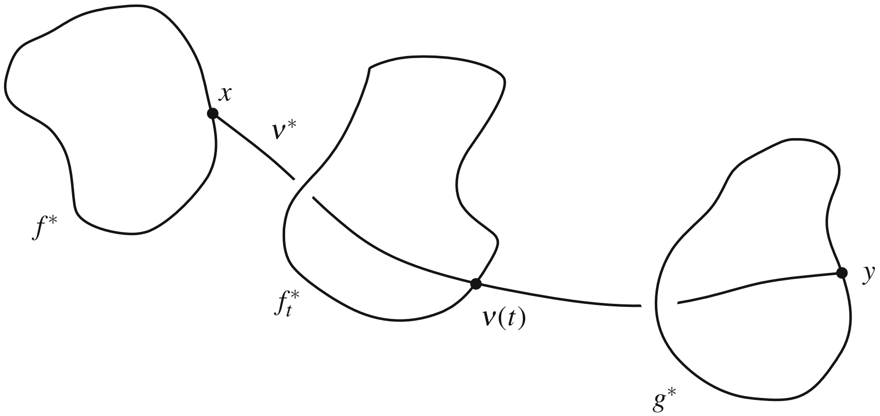
(ii) $H(0, t) = H(1, t)$ $(t \in I)$.

Note that the paths h_t, where $h_t(s) = H(s, t)$, are closed but are not required to have the same base point; the path $t \longmapsto h_t(0) : I \to X$ is not required to be a constant map. A simple example of a free homotopy occurs when the base point of a closed path is 'shifted' to another point on its track: see Exercise 2.5.30/4.

Theorem 2.5.17 *Let f and g be closed paths in a metric space X such that f is freely homotopic to g under a homotopy $F : I \times I \to X$. Let ν be the path in X from $f(0)$ to $g(0)$ defined by $\nu(s) = F(0, s)$ $(s \in I)$. Then, with $\widehat{\nu}$ given by $\widehat{\nu}(s) = \nu(1 - s)$ $(s \in I)$,*

$$f \sim \nu * g * \widehat{\nu}.$$

Proof Let $x = f(0)$, $y = g(0)$ so that ν is a path joining x to y and $\widehat{\nu}$ is its reverse: the figure below is a guide.



Recall that

$$(\nu * g * \widehat{\nu})(s) = \begin{cases} \nu(3s), & 0 \leq s \leq 1/3, \\ g(3s - 1), & 1/3 \leq s \leq 2/3, \\ \nu(3(1 - s)), & 2/3 \leq s \leq 1. \end{cases}$$

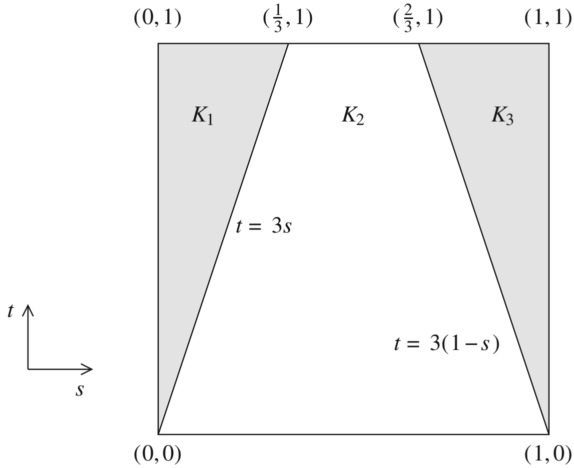
For fixed $t \in I$, consider the path γ_t given by

$$\gamma_t(s) = \begin{cases} \nu(3s), & 0 \leq s \leq t/3, \\ f_t((3 - 2t)^{-1}(3s - t)), & t/3 \leq s \leq 1 - t/3, \\ \nu(3(1 - s)), & 1 - t/3 \leq s \leq 1, \end{cases}$$

where f_t is that path such that $f_t(s) = F(s, t)$. Loosely speaking, γ_t proceeds from x to $\nu(t)$, circuits the track of f_t and then retraces its steps back from $\nu(t)$ to x. We show that $\{\gamma_t : t \in I\}$ determines a suitable homotopy. Define $G : I \times I \to X$ by

$$G(s, t) = \begin{cases} F(0, 3s), & 0 \leq s \leq t/3, \\ F\left((3 - 2t)^{-1}(3s - t), t\right), & t/3 \leq s \leq 1 - t/3, \\ F(0, 3(1 - s)), & 1 - t/3 \leq s \leq 1. \end{cases}$$

Let K_1, K_2, K_3 be the subsets of $I \times I$ defined by the inequalities $0 \le s \le t/3$, $t/3 \le s \le 1 - t/3$ and $1 - t/3 \le s \le 1$ respectively and indicated below.



It is plain that each of the following maps is continuous:
$(s,t) \longmapsto (0,3s) \longmapsto F(0,3s) : K_1 \to I \times I \to X$,
$(s,t) \longmapsto \left((3-2t)^{-1}(3s-t), t\right) \longmapsto F\left((3-2t)^{-1}(3s-t), t\right) : K_2 \to I \times I \to X$
and
$(s,t) \longmapsto (0, 3(1-s)) \longmapsto F(0,3(1-s)) : K_3 \to I \times I \to X$.
Thus each $G\,|_{K_i}$ is continuous. Since G is consistently defined on the line segments $K_1 \cap K_2$ and $K_2 \cap K_3$, and each K_i is closed, it follows from the glueing lemma that G is continuous. Now

$$G(s,0) = F(s,0) = f(s) \ (s \in I),$$

$$\begin{aligned} G(s,1) &= \begin{cases} F(0,3s) = \nu(3s), & 0 \le s \le 1/3, \\ F(3s-1,1) = g(3s-1), & 1/3 \le s \le 2/3, \\ F(0,3(1-s)) = \nu(3(1-s)), & 2/3 \le s \le 1, \end{cases} \\ &= (\nu * g * \widehat{\nu})(s) \ (s \in I) \end{aligned}$$

and

$$G(0,t) = x = G(1,t) \ (t \in I).$$

Thus $f \sim \nu * (g * \widehat{\nu})$, as required. □

Theorem 2.5.18 *Let X be a metric space and let f be a closed path in X with base point x. Then $f \sim e_x$ if, and only if, f is freely homotopic to a constant path in X.*

Proof If $f \sim e_x$ the result is obvious. Conversely, suppose that for some $y \in X$, f is freely homotopic, under a homotopy $F : I \times I \to X$, to the constant path e_y. Let ν

be the path in X from x to y given by $\nu(s) = F(0, s)$ $(s \in I)$. Then

$$\begin{aligned} f &\sim \nu * e_y * \widehat{\nu} \text{ (by Theorem 2.5.17)} \\ &\sim \nu * (e_y * \widehat{\nu}) \text{ (by Theorem 2.5.11)} \\ &\sim \nu * \widehat{\nu} \text{ (by Theorems 2.5.12 and 2.5.9)} \\ &\sim e_x \text{ (by Theorem 2.5.13).} \end{aligned}$$

Hence by Theorem 2.5.5, $f \sim e_x$. □

Definition 2.5.19 A closed path in a metric space X is said to be **null-homotopic in** X if it is freely homotopic to a constant path in X. A metric space X is said to be **simply-connected** if it is path-connected and each closed path in X is null-homotopic in X.

Remark 2.5.20 By Theorem 2.5.18, X is simply-connected if it is path-connected and $f \sim e_{f(0)}$ for each closed path f in X. Intuitively, a simply-connected space may be viewed as one within which each pair of points can be joined by a path and each closed path is continuously shrinkable to a point. No closed path can 'encompass a hole' in the space.

Example 2.5.21

(i) Let K be a subset of a metric space X. Suppose that X is also a linear space and that K is convex, so that $tx + (1 - t)y \in K$ whenever $x, y \in K$ and $t \in I$. Then K is simply-connected: its convexity implies that it is path-connected; moreover, if γ_0 is any closed path in K and $z \in K$, then γ_0 is freely homotopic to the constant path γ_1, where $\gamma_1(s) = z$ for all $s \in I$, under the homotopy $H : I \times I \to K$ defined by $H(s, t) = (1 - t)\gamma_0(s) + tz$. Hence each ball in $\mathbf{R}^n$ is simply-connected.

(ii) We shall see in Chap. 3, once the notion of winding number has been developed, that neither a circle nor an annulus in $\mathbf{R}^2$ is simply-connected.

The next two theorems reinforce the definition above and have application in the chapter to follow.

Theorem 2.5.22 *Let x and y be points in a simply-connected metric space X and let f, g be paths in X which join x to y. Then $f \sim g$.*

Proof Let paths e_x, e_y and $\widehat{g}$ be given by $e_x(s) = x$, $e_y(s) = y$ and $\widehat{g}(s) = g(1 - s)$. Note that $\widehat{g} * f$ is a closed path with base point y and, since X is simply connected, $\widehat{g} * f \sim e_y$. Since the relation $\sim$ is transitive, the steps below yield the result:

$$\begin{aligned} f &\sim e_x * f \text{ (by Theorem 2.5.12)} \\ &\sim (g * \widehat{g}) * f \text{ (by Theorems 2.5.9 and 2.5.13)} \\ &\sim g * (\widehat{g} * f) \text{ (by Theorem 2.5.11)} \\ &\sim g * e_y \text{ (by Theorem 2.5.9)} \\ &\sim g \text{ (by Theorem 2.5.12).} \end{aligned}$$

□

Theorem 2.5.23 *Let X, Y be metric spaces and let $X \times Y$ be endowed with the usual metric (see Example 2.1.2 (ix)). Then $X \times Y$ is simply-connected if, and only if, both X and Y are simply-connected.*

Proof Suppose that X and Y are simply-connected and let $\gamma = (\gamma_1, \gamma_2)$ be a closed path in $X \times Y$. Then γ_1 and γ_2 are closed paths which are null-homotopic in X and Y respectively. Let maps $F_1 : I \times I \to X$ and $F_2 : I \times I \to Y$ establish these homotopies. Then the map $F : I \times I \to X \times Y$ given by $F(s, t) = (F_1(s, t), F_2(s, t))$ shows that γ is null-homotopic in $X \times Y$. Since, by Theorem 2.4.31, $X \times Y$ is path-connected it follows that $X \times Y$ is simply-connected.

Conversely, suppose that $X \times Y$ is simply-connected. Elementary considerations show that X and Y are path-connected. Let γ_1 and γ_2 be closed paths in X and Y respectively and define $\gamma : I \to X \times Y$ by $\gamma(t) = (\gamma_1(t), \gamma_2(t))$. Then γ is a closed path which is null-homotopic in $X \times Y$ under a homotopy $H = (H_1, H_2)$, say. Since the maps H_1 and H_2 are themselves homotopies which, respectively, establish that γ_1 is null-homotopic in X and γ_2 is null-homotopic in Y, the spaces X and Y are simply-connected. □

2.5.2 The Fundamental Group

Definition 2.5.24 Let X be a metric space and $x \in X$. Let $\mathscr{L}(x)$ denote the family of all closed paths in X with base point x :

$$\mathscr{L}(x) = \{f \in C(I, X) : f(0) = f(1) = x\}.$$

By Theorem 2.5.5, the relation $\sim$ is an equivalence relation in $C(I, X)$ and therefore in $\mathscr{L}(x)$. For $f \in \mathscr{L}(x)$, let $\langle f \rangle$ denote the equivalence class associated with f :

$$\langle f \rangle = \{g \in \mathscr{L}(x) : g \sim f\}.$$

The set

$$\pi(X, x) = \{\langle f \rangle : f \in \mathscr{L}(x)\}$$

equipped with the product defined by

$$\langle f \rangle \langle g \rangle = \langle f * g \rangle$$

is called the **fundamental group of X at the base point** x.

We must justify this terminology by showing that $\pi(X, x)$ is indeed a group.

(i) Theorem 2.5.9 shows that the product of equivalence classes is well-defined:

$$\langle f\rangle = \left\langle f'\right\rangle,\ \langle g\rangle = \left\langle g'\right\rangle \ \Rightarrow f \sim f',\ g \sim g' \Rightarrow f * g \sim f' * g' \Rightarrow \langle f * g\rangle$$
$$= \left\langle f' * g'\right\rangle.$$

(ii) By Theorem 2.5.11, the product of equivalence classes is associative:

$$(\langle f\rangle \langle g\rangle) \langle h\rangle = \langle f * g\rangle \langle h\rangle = \langle (f * g) * h\rangle = \langle f * (g * h)\rangle = \langle f\rangle \langle g * h\rangle$$
$$= \langle f\rangle (\langle g\rangle \langle h\rangle).$$

(iii) Theorem 2.5.12 confirms that $\langle e_x\rangle$ is the identity:

$$\langle e_x\rangle \langle f\rangle = \langle e_x * f\rangle = \langle f\rangle = \langle f * e_x\rangle = \langle f\rangle \langle e_x\rangle .$$

(iv) Theorem 2.5.13 shows that given $\langle f\rangle$ in $\pi(X, x)$, its inverse $\langle f\rangle^{-1} = \left\langle \widehat{f}\right\rangle$:

$$\langle f\rangle \left\langle \widehat{f}\right\rangle = \left\langle f * \widehat{f}\right\rangle = \langle e_x\rangle = \left\langle \widehat{f} * f\right\rangle = \left\langle \widehat{f}\right\rangle \langle f\rangle .$$

Given distinct points x and y in X, it is natural to ask whether there is any relationship between $\pi(X, x)$ and $\pi(X, y)$. It turns out that one exists if x and y can be joined by a path in X.

Theorem 2.5.25 *Let x and y be points in a metric space X and let α be a path in X such that $\alpha(0) = x$, $\alpha(1) = y$. Then $\pi(X, x)$ and $\pi(X, y)$ are isomorphic.*

Proof As usual, let $\widehat{\alpha}(s) = \alpha(1 - s)$ $(s \in I)$. Using the notation of Definition 2.5.24, note that if $f \in \mathscr{L}(x)$, then $\widehat{\alpha} * f * \alpha \in \mathscr{L}(y)$. Consider the map

$$\phi_\alpha : \pi(X, x) \to \pi(X, y)$$

defined (see (i), below) by

$$\phi_\alpha(\langle f\rangle) = \langle \widehat{\alpha} * f * \alpha\rangle .$$

Routine use of Theorems 2.5.9 to 2.5.13 shows that
(i) for all $f, g \in \mathscr{L}(x)$,

$$\langle f\rangle = \langle g\rangle \quad \Leftrightarrow \quad \phi_\alpha(\langle f\rangle) = \phi_\alpha(\langle g\rangle);$$

(ii) for all $u \in \mathscr{L}(y)$,

$$\phi_\alpha\left(\langle \alpha * u * \widehat{\alpha}\rangle\right) = \langle u\rangle ;$$

(iii) for all $f, g \in \mathscr{L}(x)$,

$$\phi_\alpha(\langle f\rangle)\phi_\alpha(\langle g\rangle) = \phi_\alpha(\langle f\rangle \langle g\rangle).$$

Detailed proof of (i) to (iii) is left to the reader, but by way of illustration of the procedures to be adopted we indicate how to deal with (iii). For all $f, g \in \mathscr{L}(x)$,

$$\begin{aligned}
\phi_\alpha(\langle f\rangle)\phi_\alpha(\langle g\rangle) &= \langle(\widehat{\alpha} * f * \alpha) * (\widehat{\alpha} * g * \alpha)\rangle = \langle\widehat{\alpha} * f * \alpha * \widehat{\alpha} * g * \alpha\rangle \\
&= \langle(\widehat{\alpha} * f) * (\alpha * \widehat{\alpha} * g * \alpha)\rangle \\
&= \langle(\widehat{\alpha} * f) * ((\alpha * \widehat{\alpha}) * (g * \alpha))\rangle \\
&= \langle(\widehat{\alpha} * f) * (e_x * (g * \alpha))\rangle = \langle(\widehat{\alpha} * f) * (g * \alpha)\rangle \\
&= \langle\widehat{\alpha} * ((f * g) * \alpha)\rangle = \langle\widehat{\alpha} * (f * g) * \alpha\rangle = \phi_\alpha(\langle f\rangle\,\langle g\rangle).
\end{aligned}$$

Statements (i) and (ii) show that ϕ_α is well-defined and bijective; (iii) shows that it is a homomorphism. Hence $\pi(X, x)$ and $\pi(X, y)$ are isomorphic groups. □

This theorem has immediate corollaries.

Corollary 2.5.26 *Let x and y belong to a path-connected metric space X. Then $\pi(X, x)$ and $\pi(X, y)$ are isomorphic.*

Note that different paths between x and y may generate different isomorphisms.

Corollary 2.5.27 *A metric space X is simply-connected if, and only if, it is path-connected and $\pi(X, x) = \{\langle e_x\rangle\}$ for some (and thus each) $x \in X$.*

To conclude this section, we show that fundamental groups at two points, one from each of two homeomorphic, path-connected metric spaces, are isomorphic. The next result is key in this: it does not require the hypothesis of path-connectedness.

Theorem 2.5.28 *Let X and Y be homeomorphic metric spaces. Let $x \in X$ and suppose that $\psi : X \to Y$ is a homeomorphism. Then $\pi(X, x)$ and $\pi(Y, \psi(x))$ are isomorphic groups.*

Proof With the notation of Definition 2.5.24, if $f, g \in \mathscr{L}(x)$, then evidently $\psi \circ f$, $\psi \circ g \in \mathscr{L}(\psi(x))$. Further, use of Theorem 2.5.7 shows that

$$f \sim g \text{ in } \mathscr{L}(x) \;\Leftrightarrow\; \psi \circ f \sim \psi \circ g \text{ in } \mathscr{L}(\psi(x)). \tag{2.5.1}$$

Consider the map $\Psi : \pi(X, x) \to \pi(Y, \psi(x))$ given by

$$\Psi(\langle f\rangle) = \langle\psi \circ f\rangle \; (f \in \mathscr{L}(x)).$$

Because of (2.5.1), the map Ψ is well-defined and injective; it is surjective since $\Psi\left(\langle\psi^{-1} \circ u\rangle\right) = \langle u\rangle$ for each $u \in \mathscr{L}(\psi(x))$; moreover, it is a homomorphism as

$$\Psi(\langle f\rangle)\Psi(\langle g\rangle) = \langle(\psi \circ f) * (\psi \circ g)\rangle = \langle\psi \circ (f * g)\rangle = \Psi(\langle f * g\rangle) = \Psi(\langle f\rangle\,\langle g\rangle)$$

whenever $f, g \in \mathscr{L}(x)$. Thus Ψ is a group isomorphism and $\pi(X, x)$ and $\pi(Y, \psi(x))$ are isomorphic groups. □

Corollary 2.5.29 *Let X and Y be homeomorphic metric spaces, each of which is path-connected. Then, for arbitrary choice of $x \in X$ and $y \in Y$, the groups $\pi(X, x)$ and $\pi(Y, y)$ are isomorphic.*

Proof The result follows from Theorem 2.5.28 and Corollary 2.5.26. □

The message of the corollary is that homeomorphic, path-connected spaces give rise to isomorphic fundamental groups.

Exercise 2.5.30

1. Let S^n be the unit sphere in $\mathbf{R}^{n+1}$ (see Example 2.4.21 (iv)), let $f : S^n \to S^n$ be continuous, and suppose that, for all $x \in S^n$, $f(x) \neq -x$. Show that $f \simeq \mathrm{id}_{S^n}$, where id_{S^n} is the identity map on S^n. [Consider the map $H : S^n \times I \to S^n$ defined by
$$H(x, t) = \frac{(1-t)f(x) + tx}{\|(1-t)f(x) + tx\|},$$
where $\|u\| = \left(\sum_{j=1}^{n+1} u_j^2\right)^{1/2}$ for $u = (u_1, \ldots, u_{n+1}) \in \mathbf{R}^{n+1}$.]
2. Let x and y be points in a metric space X, and let $\mu, \nu : I \to X$ be paths in X from x to y. Show that $\mu \sim \nu$ if, and only if, $\mu * \widehat{\nu} \sim e_x$.
3. Give examples of closed paths f, g in $\mathbf{R}^2$ such that $(f * f) * f \neq f * (f * f)$ and $(g * g) * g = g * (g * g)$.
4. Let f be a closed path in a metric space X; let $a \in I$ and define $g : I \to X$ by
$$g(s) = \begin{cases} f(s+a) & \text{if } 0 \leq s \leq 1-a, \\ f(a+s-1) & \text{if } 1-a \leq s \leq 1. \end{cases}$$
Show that g is a closed path in X, that $g^* = f^*$ and that $H : I \times I \to X$ defined by
$$H(s, t) = \begin{cases} f(s+ta) & \text{if } 0 \leq s \leq 1-ta, \\ f(ta+s-1) & \text{if } 1-ta \leq s \leq 1 \end{cases}$$
establishes a free homotopy between f and g.
5. Generalise Example 2.5.2: let X be a metric space, Y be a subspace of $\mathbf{R}^n$ (a non-empty subset of $\mathbf{R}^n$ endowed with the metric inherited from $\mathbf{R}^n$, not to be confused with a linear subspace) and $f_0, f_1 : X \to Y$ be continuous maps such that, for all $(x, t) \in X \times I$, $(1-t)f_0(x) + tf_1(x) \in Y$. Show that $f_0 \simeq f_1$.
6. Two metric spaces X and Y are said to be **homotopy-equivalent** (written $X \simeq Y$) if there exist continuous maps $f : X \to Y$ and $g : Y \to X$ such that $g \circ f \simeq id_X$ and $f \circ g \simeq id_Y$, where $id_X : X \to X$ and $id_Y : Y \to Y$ are the identity maps. Prove that homotopy-equivalence is an equivalence relation on the family of all metric spaces. Note that homeomorphic spaces are homotopy-equivalent; also, as illustrated below, the converse need not hold.
7. (i) Let X and Y be the subspaces of $\mathbf{R}^2$ given by $X = S^1$ and $Y = S^1 \cup \{(x, 0) : 1 \leq x \leq 2\}$. Prove that X and Y are homotopy-equivalent but not homeomorphic.

[Hint: consider maps $f : X \to Y$, $g : Y \to X$ defined respectively by $f(x) = x$ if $x \in X$, $g(y) = y$ if $y \in S^1$, $g(y) = (1, 0)$ if $y \in Y \backslash S^1$.]
(ii) Let X and Y be subspaces of $\mathbf{R}^2$ given by $X = S^1$ and $Y = \mathbf{R}^2 \backslash \{0\}$. Show that X and Y are homotopy-equivalent but not homeomorphic. [Hint: consider the map $f : X \to Y$ given by $f(x) = x$, and the map $g : Y \to X$ defined by $g(y) = |y|^{-1}\, y$, where $|y| = \left(y_1^2 + y_2^2\right)^{1/2}$ for $y = (y_1, y_2) \in \mathbf{R}^2$.]

8. A metric space X is called **contractible** if the identity map $id : X \to X$ is homotopic to a constant map. Prove that X is contractible if, and only if, X is homotopy-equivalent to a space consisting of a single point. Show that every convex, non-empty subset of $\mathbf{R}^n$ is contractible.

Springer
http://www.springer.com/978-3-319-06208-2

From Real to Complex Analysis
Dyer, R.H.; Edmunds, D.E.
2014, X, 332 p. 13 illus., Softcover
ISBN: 978-3-319-06208-2